MASTER EN SCIENCES INFORMATIQUES

L'aide à la validation des données

Fruytier, Christophe

Award date:
2009

Awarding institution:
Universite de Namur

Facultés Universitaires
Notre-Dame de la Paix, Namur
Institut d'Informatique.
Année académique 2008-2009

L'aide à la validation des données

Christophe Fruytier

Mémoire présenté en vue de l'obtention du grade de licencié en informatique.

Résumé

Ce document est consacré à un problème important économiquement mais mal maîtrisé pour l'instant.

Avec l'informatisation croissante, les entreprises disposent à l'heure actuelle de grandes quantités de données. Celles-ci sont devenues au fil du temps la base de leurs décisions stratégiques. Néanmoins, ces données comportent en général un grand nombre d'erreurs. De ce fait, il est risqué de manipuler ces données tant qu'elles n'ont pas été vérifiées et nettoyées (Data Cleaning). La validation et gestion de ces données par les départements métiers comme par les départements informatiques est aujourd'hui une nécessité pour atteindre un niveau de qualité de données satisfaisant.

Ce mémoire présente les concepts et techniques de base de l'aide à la validation des données et propose des solutions concrètes pour la mise en place d'une telle aide dans le cadre de grandes entreprises.

Abstract

This document aims to present a problem economically important but not yet under control.

With the increasing use of computers, companies host today a large amount of data. These ones are today the heart of strategic decisions. However, these data are in most cases dirty. Therefore, manipulating these ones without verifications and cleaning is risky. The validation and management of data by business and IT departments together is today a necessity to reach an acceptable level of quality.

This document presents the basic concepts and techniques of data validation help and proposes concrete solutions to sustain data validation in the framework of large companies

Remerciements

Je tiens à adresser mes plus sincères remerciements à mon directeur de mémoire, Monsieur Hainaut, pour m'avoir fait part de ses conseils avisés, pour son suivi et pour m'avoir accordé sa confiance lors de l'élaboration de ce mémoire.

J'adresse également toute ma gratitude à Florence, à Anne-Catherine, à mes parents et amis qui m'ont soutenu et aidé durant la préparation du présent mémoire.

Table des matières

Introduction

Le lundi 18 mai 2009, le journal *Le Soir* publiait un article intitulé « La planète menacée par l'obésité numérique » dont voici un extrait :

> *« L'inflation des données numériques poursuit sa folle croissance et même la crise financière n'a pu l'enrayer. C'est l'un des faits saillants qui ressort d'une étude du consultant IDC, qui vient d'être rendue publique. L'an dernier, la quantité de données numériques produites sur la planète a atteint 487 milliards de gigaoctets, l'équivalent de 104 milliards de DVD remplis de données. Elle n'était « que » de 281 milliards de gigaoctets en 2007.(…) L'empreinte numérique de chaque habitant de la planète laisse pantois : en 2008, elle représentait environ 78 DVD par personne. Et l'impact sur l'environnement est accablant. Toujours selon IDC, chacun des disques durs utilisés par millions dans les centres de données de la planète coûte 25 euros par an rien qu'en frais d'électricité et de refroidissement»* [Jennotte, 2009]

A la lumière de cet article, nous prenons conscience de la place centrale qu'occupent les données numériques dans la société actuelle et du coût tant financier qu'écologique que représentent ces quantités astronomiques de données. Mais celles-ci sont-elles utilisables ou même nécessaires ?

Avec le développement de disciplines tel que le marketing, les entreprises ont pris conscience de l'importance de connaître leurs consommateurs : « *Get a 360° degree view of my customer* ». L'apparition du CRM (Customer Relationship Management ou gestion de relation client) au tout début des années 2000 n'a fait qu'accélérer ce phénomène. « *Ce concept consiste à savoir cibler, à attirer et à conserver les bons clients et représente un facteur déterminant du succès de l'entreprise* » [Communauté Wikipédia, 2009].

Une des composantes principales du CRM est la connaissance du client afin de répondre de manière plus adéquate à leurs besoins, d'accroître ainsi leur satisfaction et donc de les fidéliser.
Dans un premier temps, l'absence de données clients posait problème. Ensuite, avec l'informatisation croissante de tous les secteurs de l'économie, les sociétés ont alors pu investir dans des collectes massives de données et la création de *data warehouse* ou entrepôt de données.

Paradoxalement, elles se sont rendu compte assez tardivement de l'importance de la qualité de ces données et de la nécessité de les valider afin d'écarter les données inutilisables.

Ce mémoire est donc consacré à un problème important économiquement mais mal maîtrisé pour l'instant.

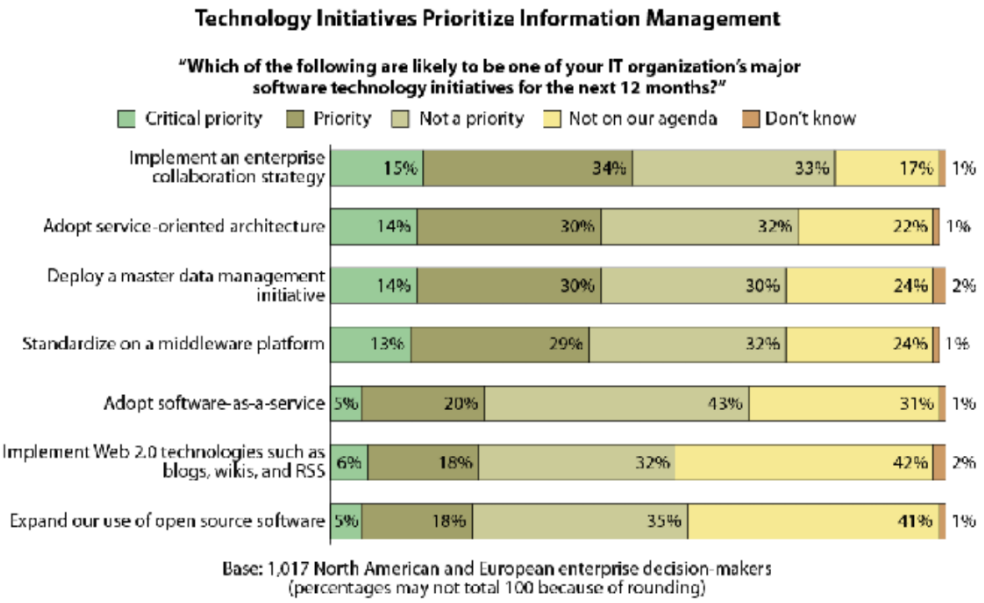


Figure 1 : Priorité des initiatives [Karel-Forrester, 2008]

Comme le montre la figure 1, la gestion des données et de leur qualité est devenue à l'heure actuelle une des principales priorités des directeurs de départements informatiques. Néanmoins, ce constat est à nuancer. Si tout le monde est conscient qu'il s'agit d'un élément crucial pour les entreprises, il n'en reste pas moins difficile d'obtenir des budgets pour de tels projets et ce pour plusieurs raisons :

1. Ces types de projet sont généralement considérés comme purement techniques car souvent transparents pour l'utilisateur.

2. Les bénéfices que peuvent en tirer les départements opérationnels sont souvent difficiles à mettre en évidence. « Pourquoi changer quelque chose qui fonctionne ? ».

3. Pour des questions de « prioritisation », ces projets sont postposés.

4. Une méconnaissance de l'état actuel des données gérées par l'entreprise ne permet pas aux dirigeants de se rendre compte de l'ampleur du problème.

5. Enfin, s'engager dans une amélioration de la qualité des données ne consiste pas à s'engager dans un projet avec une date de fin bien déterminée mais dans un programme à long terme impliquant souvent la mise en place de structures organisationnelles dédiées à ce sujet. Il s'agit de signer un contrat à durée indéterminée dont les coûts font souvent peur aux dirigeants.

Nous devons également mettre ce constat en perspective avec le contexte économique actuel. Il va de soi que celui-ci aura également un impact sur la mise en place de tels programmes. Néanmoins, cet impact ne sera pas nécessairement négatif. En effet, la crise poussera certaines entreprises à améliorer la qualité de leurs données afin justement de pouvoir réagir au mieux aux situations que nous connaissons à l'heure actuelle. Cette crise peut ou doit être perçue par les *data managers* comme une opportunité de vendre aux dirigeants de leur entreprise de tels projets.

L'apport de ce mémoire se situe principalement dans le fait qu'il ne prend pas uniquement en compte l'aspect technique de l'aide à la validation des données mais également d'autres dimensions. En effet, au-delà du fait de constater au moyen d'outils informatiques qu'une donnée est manquante, en double ou erronée, nous montrerons que ce sujet doit être considéré dans un cadre plus global. Il doit être envisagé dès la mise en place d'un projet informatique ou même organisationnel, pris en considération aussi bien par les départements métiers que par les départements IT et enfin aborder dans le cadre d'une interconnexion croissante des systèmes applicatifs de par la mondialisation. La validation des données est entièrement dépendante de son contexte d'utilisation.

Nous nous sommes rendu compte lors de recherches et expériences liées à ce mémoire que la complexité de la validation des données et de la notion de qualité ne se trouvaient pas au niveau des outils à mettre en place pour contrôler les données mais bien au niveau de l'organisation à mettre en place pour procéder à de telles validations. Nous verrons également que la partie la plus critique du processus n'est pas de la responsabilité des

départements IT mais bien des départements métiers qui doivent penser, définir et maintenir les règles de validation.

Ce mémoire abordera dans un premier temps les concepts fondamentaux de l'aide à la validation des données. Nous traiterons tout d'abord de l'objectif de l'aide à la validation des données à savoir l'amélioration de la qualité des données (chapitre 1). Ensuite, nous présenterons la définition et le processus général de l'aide à la validation des données (chapitre 2).
Dans le chapitre 3, nous exposerons les cadres d'utilisation dans lesquels l'aide à la validation constituera une aide précieuse.
Cette première partie permettra ainsi au lecteur de mieux comprendre la problématique, le processus ainsi que le contexte du sujet abordé.

La seconde partie du mémoire présentera une description des différents types d'erreurs pouvant être rencontrés au niveau de la donnée (chapitre 4) ainsi que les solutions que l'informatique peut apporter à ces problèmes, qu'elles soient techniques ou méthodologiques (chapitre 5).

La troisième partie passera en revue les trois grandes étapes du processus défini au chapitre 2. Premièrement, nous aborderons la définition des besoins en termes d'aide à la validation (chapitre 6). Deuxièmement, nous présenterons les techniques permettant de localiser et traiter les erreurs et, ce, en mettant l'accent sur un angle d'action particulier: agir sur l'utilisateur. Nous y traiterons des solutions organisationnelles qui peuvent faciliter la validation des données (chapitre 7). Enfin, nous verrons comment réaliser l'évaluation de la qualité des données obtenue (chapitre 8). Cette troisième partie a pour but de donner au lecteur des moyens concrets qui lui permettront de mettre place au travers d'une approche structurée un programme d'aide à la validation des données.

Dans la quatrième partie du mémoire, nous verrons quelle organisation mettre en place (data management) afin d'assurer la qualité des données sur le long terme (chapitre 9).

Enfin, la dernière partie sera consacrée aux conclusions de ce mémoire ainsi qu'aux trois années de recherches et d'expériences accumulées sur le sujet.

1ère Partie : Concepts fondamentaux de l'aide à la validation des données

1. La qualité des données

Afin de comprendre ce qu'est la validation des données et les processus qui y sont attachés, nous devons tout d'abord nous pencher sur la notion de qualité des données.

Nous présenterons tout d'abord l'importance de la qualité de l'information. Ensuite, nous proposerons au travers de la littérature une définition de ce qu'est la qualité des données. De cette analyse découlera la question de la qualité totale. Est-elle réalisable et quel est son coût ?

1.1 Importance et impact de la qualité des données

La qualité des données est primordiale, surtout à l'heure actuelle, où un nombre sans cesse croissant d'entreprises basent leurs décisions stratégiques sur des quantités de plus en plus grandes de données parfois collectées à prix d'or. Ainsi, l'importance de la qualité des données qui sont traitées apparaît au grand jour au travers de conséquences, tantôt loufoques, tantôt dramatiques, engendrées par certaines informations erronées. Dans certains secteurs tels que les soins de santé, la qualité des données peut devenir vitale !

Afin d'illustrer ces propos, il suffit d'ouvrir le journal ou d'allumer la télévision. Une mauvaise qualité de données peut très vite avoir des conséquences désastreuses. L'actualité de ces dernières années a été révélatrice dans ce domaine :

- « En mai 1999, durant la guerre en Bosnie, l'armée américaine a bombardé par inadvertance l'ambassade chinoise. Il est apparu que le bombardement provenait directement d'une erreur de données. Les informations associées à la localisation de la cible n'étaient tout simplement plus d'actualité. » [traduit de Redman, 2004]

- « Le cas Jessica Santillán. Afin de tenter dans un ultime recours de lui sauver la vie, les médecins lui greffèrent un bloc cœur-poumons. Malheureusement, une erreur de données entraina le rejet du greffon et le décès de la patiente. Le comité des soins de santé aux Etats-Unis estima que pas moins de 98.000 décès par an, dans de nombreux cas

dus à une mauvaise qualité de données, auraient pu être évités.» [traduit de Redman, 2004]

• La qualité des données fut également incriminée en 2000 lors des élections présidentielles aux Etats-Unis mais également dans le cadre des attaques terroristes du 11 septembre 2001. [Redman, 2004]

• Plus proche de chez nous, en 2006, l'administration des finances belge perd 900 millions d'euros. Une fois encore la qualité des données est mise en cause.

Comme nous pouvons le voir au travers de ces exemples, l'importance de la qualité des données ne nécessite pas de longues argumentations. Thomas Redman [Redman, 1995] explique en trois points pourquoi la qualité des données doit être considérée comme stratégique:

• Premièrement, la mauvaise qualité des données est **dominante**. Il s'agit d'une plaie face à laquelle aucune industrie, gouvernement ou académie n'est immunisé. Les exemples cités ci-dessus en sont la preuve. Mais il est également facile de trouver des situations plus communes. En effet, qui n'a jamais fait l'objet d'une erreur de facturation, acheté un produit en promotion au supermarché et s'apercevoir à la caisse que la réduction n'avait pas été appliquée. Les exemples sont nombreux et touchent chacun d'entre nous. Ce caractère dominant a été confirmé par de nombreuses études.

• Deuxièmement, la mauvaise qualité des données a un **impact conséquent sur l'économie** et ce à différents niveaux.

- Elle rend la satisfaction du consommateur beaucoup plus difficile à atteindre. Un bon exemple est celui de la facturation. Tout consommateur ne souhaite-t-il pas être facturé à hauteur de ce qu'il doit et pas plus ?
- Ensuite, une mauvaise qualité des données entraîne une augmentation des coûts. Si le coût réel est extrêmement difficile à évaluer, nous nous accordons à dire qu'il est très élevé. Aux Etats-Unis, The data warehousing Institute estima que le coût de la mauvaise qualité des données se rapportant aux consommateurs (nom et adresse) coûtait chaque année 611 milliards de dollars.
- Une mauvaise qualité des données amène à une diminution de la qualité de l'emploi et à des tensions. En effet, les

services en lien direct avec le consommateur font face à des gens irrités. Ce mécontentement étant causé dans de nombreux cas par des données erronées. Ceci rend le travail des opérateurs plus difficile et stressant.

- Les décisions et les stratégies à long terme basées sur des données erronées peuvent être catastrophiques pour une entreprise.
- Enfin, une mauvaise qualité des données freine les processus de *re-engineering*. Une qualité déficiente des données préexistantes dans le cadre d'une migration peut être un sérieux frein voire causer l'échec d'une migration de systèmes.

• Troisièmement, des données de qualité peuvent être un **avantage concurrentiel de premier ordre**. « *Il apparaît clairement que dans le monde des industries, des entreprises et des administrations, la qualité de l'information est l'un des enjeux financiers et compétitifs les plus importants*» [Boydens, 1998].

1.2 Définition de la qualité des données

La qualité des données est un des principaux domaines de recherche actuel en particulier dans le cadre des *data warehouse*. La littérature présente le concept de qualité des données comme un concept multidimensionnel [Wand et Wang, 1996], c'est-à-dire que l'on attend d'une donnée qu'elle remplisse un certain nombres de critères pour pouvoir dire qu'elle est de qualité ou non. Il n'y a cependant pas un seul et même point de vue sur ces dimensions. Ainsi, plusieurs dizaines de dimensions ont été présentées. Cette diversité de dimensions est principalement due au fait que la qualité des données ne peut être envisagée indépendamment du contexte d'utilisation des ces mêmes données et de ce fait est un concept aux multiples variables.

Sur base de ce constat, nous avons décidé de retenir la définition suivante :

"Data are of high quality if they are fit for their intended uses in operations, decision making and planning (based on J.M. Juran)"

[Redman, 2000]

En effet, cette définition met en avant le caractère contextuel de la qualité des données, dans le sens où celle-ci ne dépend pas de critères invariables. La qualité d'une donnée peut être perçue différemment en fonction de l'utilisateur. C'est cette position centrale de l'utilisateur et de ses besoins qui doit être mise en avant lorsqu'on aborde le sujet de la validation des données. Dans le cadre de ce travail, c'est justement ce caractère changeant de la notion de qualité qui est intéressant mais qui rend la tâche de validation des données également plus ardue. En effet, nous pouvons d'ores et déjà dire que, quelle que soit la méthode de validation, celle-ci devra prendre en compte le contexte dans lequel elle sera appliquée, ce qui nécessitera indéniablement une grande flexibilité des solutions qui devront être apportées dans ce domaine.

La définition choisie permet également d'aborder un concept plus général et de plus en plus repandu qui est celui de *trusted data*. Ce concept consiste à dire que l'utilisateur des données doit pouvoir les manipuler sans avoir aucune réserve quant à leur qualité [Karel-Forrester, 2008]. Au-delà de la notion de qualité, il faut que l'utilisateur ait confiance dans les données qu'il utilise.

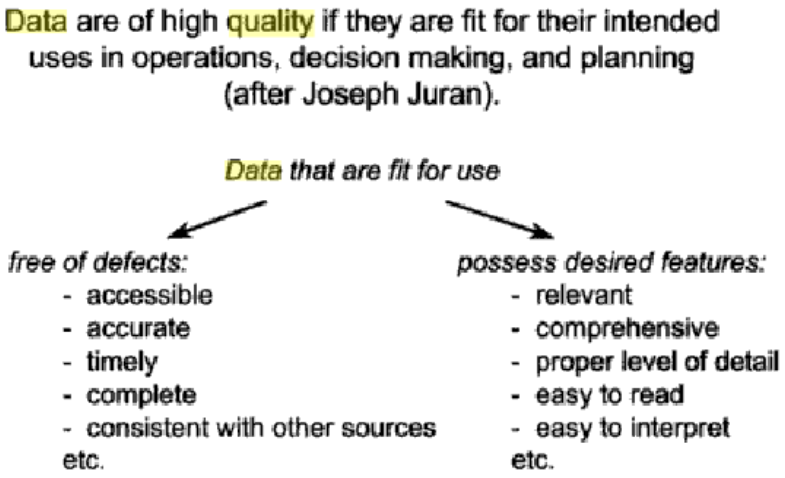


Figure 2 : définition de la qualité des données [Redman,2000]

Notons enfin, que le choix de cette définition ne diminue pas pour autant l'importance des nombreuses dimensions décrites dans la littérature. En effet, celles-ci permettent de donner un canevas aux analyses contextuelles qui doivent être diligentées dans le cadre d'un programme d'amélioration de la qualité des données. Nous reviendrons sur ce point au chapitre 6.

1.3 La qualité : l'objectif à atteindre

1.3.1 Qualité et aide à la validation des données

Le maintien ou l'amélioration du niveau de qualité des données n'est autre que l'objectif que se fixe l'aide à la validation des données.

Cette première configuration, le **maintien du niveau de qualité**, sera principalement rencontrée dans le cadre d'intégration ou de migration de données. L'objectif de la validation des données dans de tels projets n'est en effet pas toujours une amélioration de la qualité des données mais parfois uniquement de s'assurer que les données existantes ont été intégrées ou migrées sans altération. Il va de soi que cette configuration sera également rencontrée lorsque le niveau de qualité des données d'un système est jugé satisfaisant et ne nécessite donc pas d'amélioration mais bien de la maintenance. En effet, les données sont en constant changements et requièrent donc une validation continue afin de maintenir le niveau de qualité désiré.

La seconde configuration, l'**amélioration de la qualité des données**, est quant à elle majoritairement abordée par la littérature. Et ce pour deux raisons très simples : premièrement, rares sont les entreprises aujourd'hui qui peuvent se venter d'un niveau de qualité de données satisfaisant et deuxièmement, si on est capable d'améliorer la qualité de ces données, on est à fortiori capable de la maintenir. C'est principalement sous cet angle, c'est-à-dire dans un objectif d'amélioration de la qualité des données, que nous aborderons l'aide à la validation des données dans le cadre de ce mémoire.

Les plans d'amélioration de la qualité des données ne doivent pas être envisagés comme des projets temporaires. L'amélioration de la qualité des données doit être une tâche sans cesse renouvelée.

1.3.2 Qualité parfaite et *business case*

Eradiquer la totalité des erreurs au sein d'un ensemble de données est extrêmement couteux et, dans bien des cas, irréalisable voire inutile. En effet, comment identifier des erreurs dans des données clients dues, par exemple, à des omissions ou mensonges de ceux-ci. Bien sûr, nous pouvons recouper ces données avec d'autres informations mais même en utilisant ces

techniques, qui relèvent plus de l'enquête policière, nous ne serons pas toujours capable de détecter ces erreurs.

Dès lors, nous considérons qu'un niveau de qualité de 100% ne peut ou ne doit être atteint que dans des cas très spécifiques et qu'en définitive ce niveau de qualité ultime ne doit pas être un objectif en soit pour la plupart des entreprises.

Si le but n'est pas le *zero defect*, comment déterminer le niveau de qualité à atteindre ? En définitive, il convient de procéder, comme pour tout projet industriel, à une analyse coût-bénéfice. Pour ce faire, nous présentons ici les principaux coûts et bénéfices d'un programme d'amélioration de la qualité des données :

- Coûts :

Les coûts d'un tel programme sont relativement faciles à identifier. Il s'agit par exemple des coûts de développement d'outils de validation ou encore le coût de la mise en place de nouvelles procédures permettant de contrôler la qualité des données. Il faut également tenir compte d'un possible impact sur la rapidité d'exécution de certains processus métiers.

- Bénéfices :

Il faut ici distinguer deux types de bénéfices, d'une part les bénéfices directs et d'autre part les bénéfices indirects. Les bénéfices directs recouvreront les revenus directement identifiables suite à une augmentation de la qualité des données. Les bénéfices indirects sont quant à eux moins tangibles. On parlera ici de l'amélioration de l'image de l'entreprise ou encore de l'augmentation de la satisfaction du client.

Seuls les bénéfices directs seront pris en compte dans le cadre d'une analyse coûts-bénéfices. Les bénéfices indirects seront souvent considérés comme une plus-value additionnelle ou encore une marge de sécurité. Un programme d'amélioration de la qualité des données ne générant pas de bénéfices directs ne devrait pas être entrepris.

Nous présentons ci-dessous deux exemples de calculs des bénéfices directs d'un programme d'amélioration de la qualité des données.

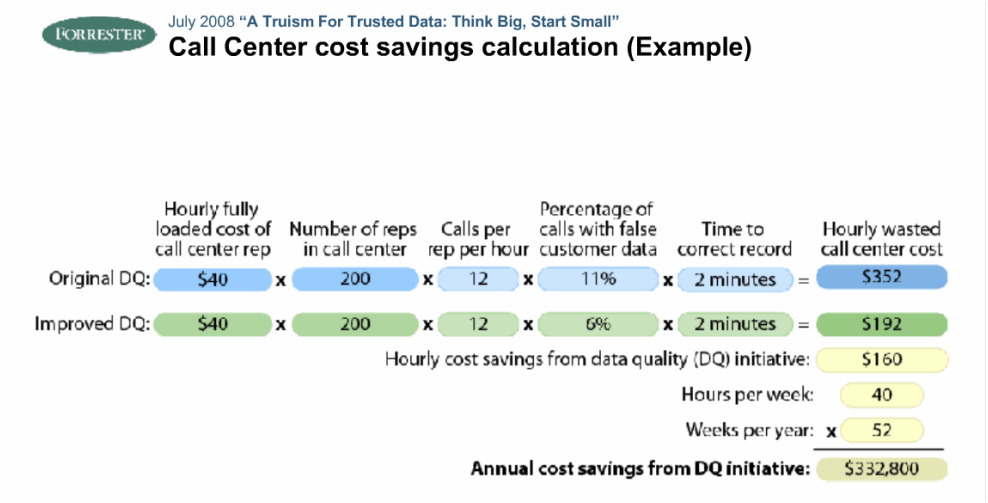


Figure 3: exemple de calculs de bénéfices [Karel-Forrester, 2008]

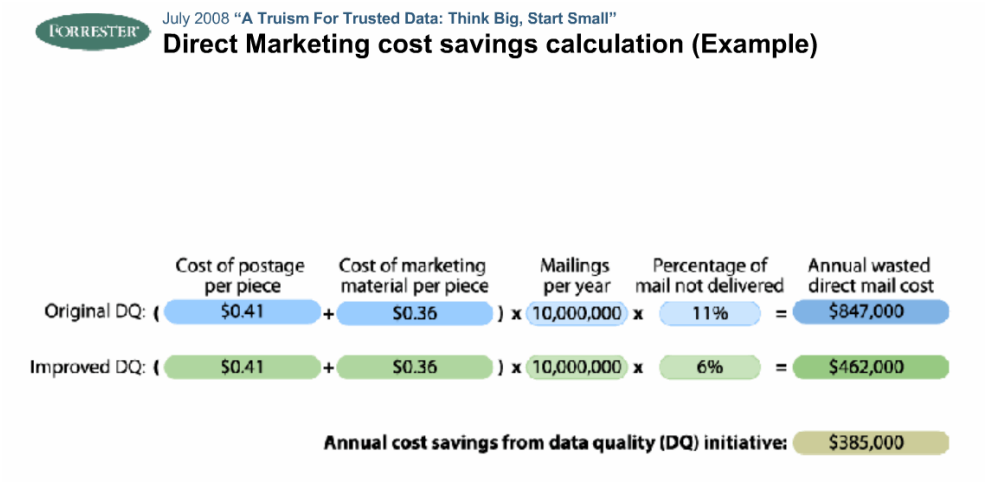


Figure 4: exemple de calculs de bénéfices [Karel-Forrester, 2008]

Sur base de cette analyse coûts-bénéfices, C. Cappielo identifie le niveau maximum de qualité de données qu'une entreprise doit atteindre comme étant le niveau où les bénéfices directs de l'amélioration du niveau de qualité équivalent aux coûts qu'ont engendré cette amélioration. Les bénéfices indirects de cette augmentation constituent quant à eux la plus-value de l'opération. [Cappiello, 2009]

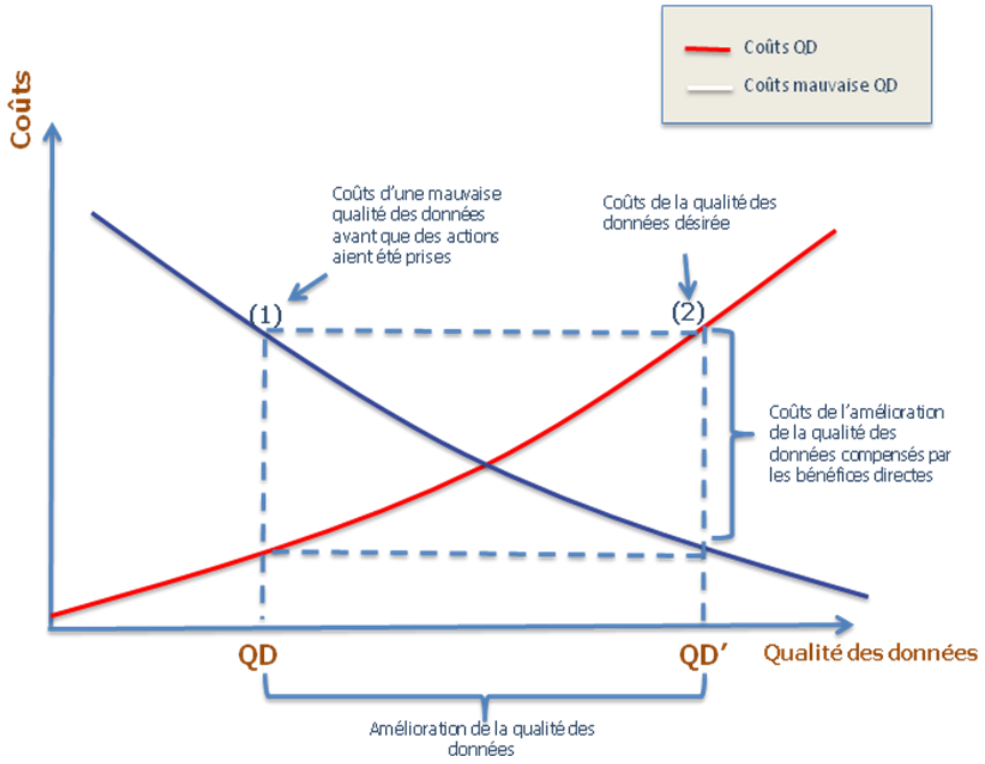


Figure 5: Analyse coût-bénéfice [adapté de Cappiello 2009]

En conclusion, il est important de maintenir l'équilibre entre les coûts nécessaires pour atteindre un niveau de qualité de données et les bénéfices que celui-ci peut engendrer. Ainsi, il convient de cibler les actions à entreprendre. Il n'est peut-être pas utile de déployer des efforts considérables pour garantir la qualité d'un champ de type « commentaire » relatif aux clients mais par contre, il est judicieux de mettre tout en œuvre pour garantir la qualité du champ « montant » relatif aux factures.

2. L'aide à la validation des données

Le concept d'aide à la validation des données n'est en soit pas directement compréhensible ou, si il l'est, il peut avoir une signification différente pour chaque lecteur. C'est pourquoi, ce chapitre a pour but d'établir une définition précise de ce qu'est l'aide à la validation des données telle qu'abordée dans ce mémoire. Nous présenterons d'abord quelle a été notre approche du sujet et nous donnerons ensuite une définition du concept traité ici.

2.1 Approche

Tout d'abord, ce mémoire se consacre à l'**aide** à la validation des données. C'est-à-dire à tous moyens informatiques mais également organisationnels capables de faciliter la validation des données. Sur base de ceci, nous avons décidé d'aborder ce sujet sous deux aspects majeurs.

Le premier consiste à présenter l'aide à la validation des données sous son aspect le plus intuitif c'est-à-dire l'aspect curatif. Celui-ci consiste à prendre comme *input* une ou plusieurs sources de données et de procéder à l'identification et au traitement des données présentant des anomalies. On ne traite donc pas ici le problème à la source mais à posteriori.

Le second aspect consiste à « dépoussiérer » la notion d'aide à la validation des données en ne l'appliquant pas uniquement à des données préexistantes mais en faisant intervenir ce processus au moment ou avant que la donnée ne soit créée. Il s'agit ici de prévenir par des moyens techniques mais surtout par des moyens organisationnels l'enregistrement de données non-valides. Dans le prolongement de ces moyens, il s'agit également de mettre en place les outils nécessaires à l'évaluation de la qualité des données tout au long de leur durée de vie. Dans cette optique, nous insisterons sur la description des besoins de l'utilisateur en termes de données ainsi que sur la documentation détaillée de ces besoins et de leur implémentation dans les systèmes.

Alors que l'approche curative s'attaque directement au problème de qualité existant, l'approche préventive s'attaque, elle, à l'origine des problèmes. La chronologie de ces approches sera dans la majorité des cas dans un premier temps une approche curative permettant ensuite de mettre

en place une aide à la validation préventive. Les deux approches ne sont et ne doivent pas être indépendantes.

Nous avons parlé de moyens techniques et organisationnels, mais qu'entend-on par ces termes? Nous utiliserons fréquemment ces termes tout au long de ce mémoire. Une définition est donc nécessaire :

- Moyens techniques :

Nous entendons par moyens techniques, toute technologie disponible permettant de faciliter la validation d'une donnée par rapport à un standard prédéfini.

- Moyens organisationnels :

Nous entendons par moyens organisationnels toute procédure, standard ou organisation permettant de faciliter la validation des données par rapport à des standards prédéfinis.

Il est clair que ces moyens ne doivent pas être envisagés en opposition mais comme des outils complémentaires.

Enfin, nous nous sommes rendu compte tout au long de nos expériences, nos recherches et nos rencontres avec des sociétés spécialisées dans le secteur de la qualité des données que ce n'était pas tant les moyens techniques (algorithme,…) qui manquaient mais plutôt la capacité des départements métiers et informatiques :

- à définir une stratégie claire en matière de qualité de données ;
- à utiliser les techniques disponibles de manière optimale ;
- et enfin, à mettre en place l'organisation nécessaire pour maintenir la qualité des données.

Au travers de notre approche, nous abordons donc le sujet de l'aide à la validation des données de manière très large en envisageant les moyens techniques et organisationnels, dans des rôles aussi bien curatifs et que préventifs et la gestion de ceux-ci. En effet, les problèmes de qualité de données sont très variés et touchent indifféremment tous les secteurs d'une entreprise. De ce fait, ils doivent être abordés de manière globale afin d'en envisager les conséquences au niveau de l'entreprise mais également afin de

mettre en place des solutions complètes en faisant appel aux connaissances de tous les départements. Nous nous inscrivons ainsi dans la lignée directe du concept de *Data Quality Management* dérivé du concept *Total Quality Management* qui a été appliqué avec succès dans le secteur des entreprises manufacturières et qui a été adapté aux données [Helfert,2001].

L'approche proposée en *data quality management* est cyclique et est décrite à la figure 6.

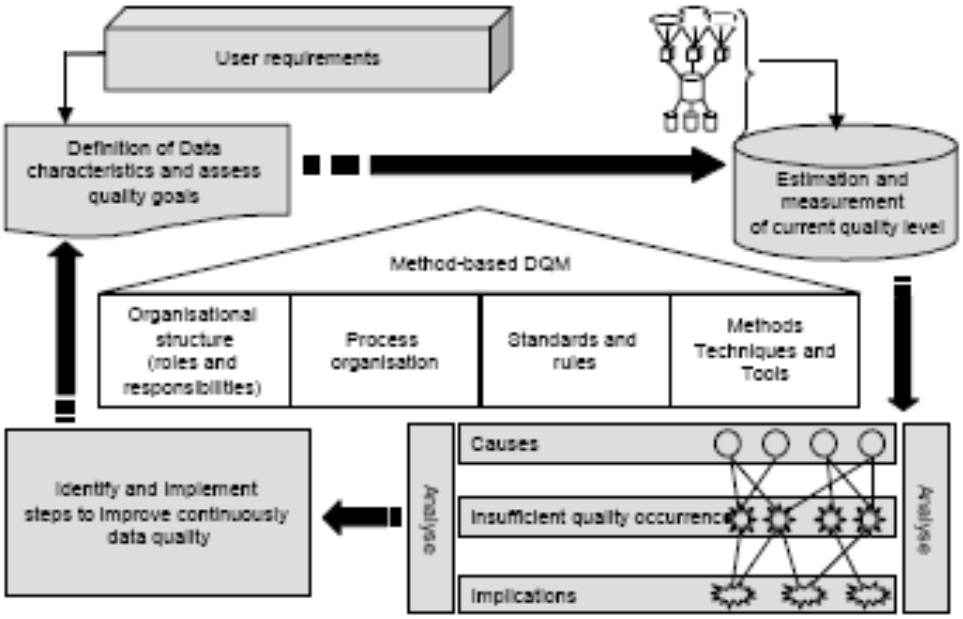


Figure 6: Method-Based Data Quality Management[Helfert, 2001]

2.2 Définition

Sur base de l'approche présentée dans la section précédente, la définition que nous proposons est la suivante :

L'aide à la validation des données englobe les moyens organisationnels et techniques qui déterminent quelles sont les données valides, facilitent la localisation des données problématiques et enfin les traitent de manière manuelle et/ou automatique afin de les conformer ou de les supprimer.

Cette définition comporte cinq points clés :

- Organisationnel et technique

Comme décrit au début de ce chapitre, la prise en considération de ces deux moyens est primordiale. Notre expérience nous a montré la nécessité d'aller au-delà des solutions techniques. En effet, celles-ci montrent assez rapidement leurs limites. Elle sont généralement très spécifiques, dédiées à des données bien précises comme les noms des clients et leurs adresses, souvent peu flexibles et nécessitant généralement l'intervention d'entreprises spécialisées généralement coûteuses. En effet, les entreprises ne disposent pas toujours des compétences nécessaires dans ce domaine particulier. Les moyens techniques doivent donc être utilisés avec parcimonie et toujours en combinaison avec des moyens organisationnels qui permettront d'en tirer le meilleur parti.

- Détermine ce qu'est une donnée valide :

La validation des données n'est autre que la confrontation d'une donnée par rapport à un standard. Une des étapes incontournables de ce processus est donc de déterminer de manière concrète les objectifs à atteindre. Il convient dans un premier temps de définir au sein de l'entreprise ce qu'on entend par « des données de qualité ». La définition et la documentation des standards et des règles de qualité qui y sont associées sont les éléments de base de la validation des données. Cette étape suppose donc une collaboration étroite entre les départements informatiques et les départements métiers. Ils doivent être élaborés ensemble et non imposés par l'un ou par l'autre comme c'est encore trop souvent le cas. Les compétences des deux bords sont nécessaires à l'élaboration de standards de qualité.

- Localise :

Une fois les contraintes établies, il s'agit de repérer les données qui les violent. La localisation des données potentiellement problématiques est le cœur du procédé et la partie la plus délicate. Néanmoins, le traitement de la donnée au moment de sa création permet une simplification à ce niveau.

- Traite :

Dans tous les cas, les données identifiées comme pouvant poser problème doivent être traitées. Ce traitement peut prendre l'aspect

d'une mise en conformité qu'elle soit effectuée par l'utilisateur ou par le programme ou par une suppression pure et simple de la donnée ou de l'enregistrement.

- Manuelle et/ou automatique:

Dans la mesure où nous parlons d'aide, il est logique que le processus soit au moins partiellement automatisé. L'automatisation complète, elle, est plus problématique. En effet, il est difficile de définir de manière automatique toutes les contraintes que doivent respecter les données pour être valides. Ceci implique que dans bien des cas une validation finale devra encore être effectuée manuellement. Néanmoins, l'objectif de l'aide à la validation des données est de réduire au maximum cette intervention manuelle.

2.3 Processus de mise en place de l'aide à la validation des données

D'après la définition de l'aide à la validation des données proposée précédemment, nous pouvons établir le processus de mise en place de celle-ci comme suit :

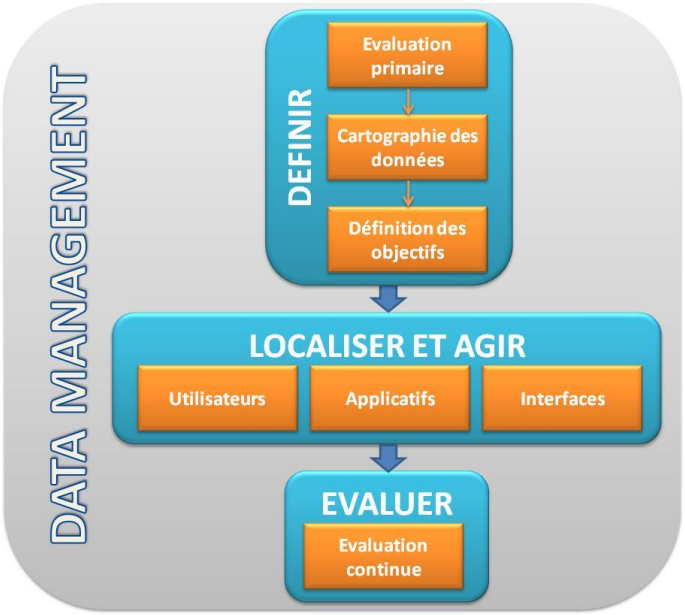


Figure 7: Processus de mise en place de l'aide à la validation des données

La troisième partie de ce mémoire aura pour objectif de décrire trois des étapes de ce processus à savoir comment définir les besoins en termes d'aide à la validation, comment localiser des données potentiellement problématiques et agir sur celles-ci, et comment évaluer les résultats de l'aide à la validation des données. Enfin, nous aborderons dans la quatrième partie de ce mémoire la gestion journalière des données et des aides à la validation mises en place au travers du *data management*.

2.4 Hypothèses

Nous avons choisi de fixer deux hypothèses dans le cadre de ce mémoire. La première consiste à ne considérer la problématique de l'aide à la validation des données que dans le cadre d'entreprises de grande taille et la suivante consiste à fixer la notion de donnée.

2.4.1 Hypothèse 1 : Les entreprises de grande taille

Nous avons décidé d'aborder le sujet de ce mémoire dans le cadre des grandes entreprises et ce pour deux raisons.

La première est que les entreprises de grande taille font généralement face à de grandes quantités de données et souvent gérées au moyen de CRM (Customer Relationship Management) du moins pour ce qui est des données clients. Cette masse de données nécessite une gestion très avancée au niveau de la validation de celles-ci.

La seconde raison est la séparation des tâches au sein de ces entreprises. Elle rend l'interaction difficile entre l'utilisateur final et le programmeur qui met en place les outils de validation. La résolution de ce problème nécessite une organisation dédiée à la gestion des données qui rend possible l'interaction entre ces deux pôles.

Ces deux éléments rendent l'aide à la validation des données absolument nécessaire dans ces entreprises. C'est pourquoi nous souhaitions axer notre réflexion sur celles-ci.

2.4.2 Hypothèse 2 : Les données informatiques

Il est nécessaire de préciser la notion de données utilisées dans ce mémoire. Nous ne traitons ici que les données stockées sur supports informatiques. Nous n'aborderons pas le problème de la validation des données stockées sur d'autres supports tel que le papier. Ces données n'offrant pas ou peu de possibilités de traitement informatisé.

3. Les cas d'utilisation

Afin de mieux comprendre les *challenges* auxquels nous devons faire face dans le cadre de l'aide à la validation des données, ce chapitre a pour objectif de décrire dans quels contextes celle-ci peut intervenir. Nous présentons ici quatre cas d'utilisation durant lesquels la validation des données doit être abordée.

3.1 La validation continue

L'optimisation continue consiste à maintenir et à améliorer continuellement la qualité des données au sein de systèmes de gestion de bases de données (SGBD) existants. Au fil du temps, si aucun système de validation des données n'est mis en place, la qualité des données décroit, la base de données accumulant les erreurs. Des imperfections dans la saisie des données, des erreurs dans les programmes en amont du SGBD,... entraînent un nombre conséquent de données erronées. De plus, les systèmes applicatifs sont souvent trop peu flexibles et donc, une fois implémentés, ne sont pas facilement modifiables. Le contexte dans lequel les données sont utilisées quant à lui change. Ce phénomène entraîne des déviations comme l'utilisation de champs pour d'autres valeurs que celles définies au départ ou encore entraine l'obsolescence de certaines informations récoltées.

Afin d'éviter ces désagréments, les entreprises souhaitent mettre en place un processus continu de validation des données. Nous retrouvons dans ce processus un cas d'utilisation de l'aide à la validation des données où l'aspect curatif et l'aspect préventif travaillent de concert.

L'aspect curatif prendra généralement la forme de programmes *batch* tournant durant la nuit de manière récurrente afin d'éviter les problèmes de performance souvent centraux dans les programmes d'aide à la validation des données. Ceux-ci cibleront particulièrement les validations trop coûteuses en termes de performance et les erreurs qui ne peuvent pas être contrôlées avant l'enregistrement de la donnée tels que des montants dont la qualité doit être contrôlée sur base de statistiques mensuelles.

L'aspect préventif dans le cadre de la validation continue lui envisagera le contrôle des données au moment de leur saisie ou de leur réception tels que les contrôles de formats de messages par exemple. On retrouvera ici des contrôles instantanés et peu exigeants en termes de performance. Ces contrôles peuvent également prendre l'aspect de procédures. Cet aspect préventif est souvent dénommé *Data Quality Firewall*.

3.2 L'intégration

Avec la mondialisation du commerce, on assiste à une intégration croissante des systèmes informatiques. Les fusions d'entreprises, les synergies, les partages et les transmissions de données, la standardisation croissante des flux informatiques sont autant de facteurs qui amènent les entreprises à intégrer dans leurs systèmes de nouveaux flux de données aux cotés de ceux déjà existants. Ces nouvelles données ne doivent pourtant pas altérer la qualité des données déjà établie. Il est donc important pour ces entreprises de s'assurer de la qualité des nouvelles données qui seront intégrées dans leurs systèmes. Celles-ci feront donc l'objet de contrôles et dans certains cas de manipulations afin de garantir des données de qualité.

Non seulement la validation préalable des données présentes dans les systèmes à intégrer est primordiale mais un défi encore plus grand est d'intégrer des données de structures différentes représentant un même objet comme le présente la figure 8 dans le cas d'un client.

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Figure 8 : Deux représentations différentes d'un client [Erhard Rahm et Hong Hai Do]

Cette figure représente les erreurs que l'on peut trouver au sein d'une source tels qu'une adresse incomplète ou encore un doublon. Mais elle met également en évidence la complexité des intégrations de données. On remarque ici que la structure des tables « Customer » et « Client » est fort

différente. De plus, dans cet exemple, les attributs ont des valeurs différentes pour représenter un même fait. C'est le cas de l'attribut « Sex » de la source 1 qui utilisera les valeurs « 1 » et « 2 » pour distinguer un homme et une femme alors que la source 2, elle, utilisera les valeurs « M » et « F ». Dans ce cas précis, on constate que la nomenclature même des attributs est différente et peut également accroître la confusion. Ensuite, beaucoup d'attributs repris dans cet exemple sont très peu structurants (« address », « Phone/Fax ») rendant l'intégration d'autant plus compliquée. Pour finir, une des questions les plus cruciales auxquelles devra répondre le processus d'intégration est : Quels sont les « Smith » de la source 1 qui correspondent aux « Smith » de la source 2?

Ainsi, le premier défi à relever dans le cadre d'une intégration de données se trouve donc au niveau des schémas de données. Dans la majorité des cas, les différentes sources auront une façon distincte de représenter la réalité. Cette complexité apparaît à deux niveaux :

Le premier est celui **des entités et de leurs attributs**. Une entité est-elle caractérisée de la même façon dans le SGBD 1 et le SGBD 2. Les attributs utilisés sont-ils semblables ? (adresse structurée dans un cas et non structurée dans l'autre). Une entité du SGBD 1 est-elle également représentée par une entité dans le SGBD 2? Si oui, par une ou plusieurs entités ?

Le deuxième niveau de complexité se situe **au niveau des relations entre les entités**. Il s'agit certainement de l'aspect le plus compliqué à gérer dans le cadre d'une intégration. Un schéma peut autoriser des relations entre certaines entités alors qu'un autre les limite ou les interdit.

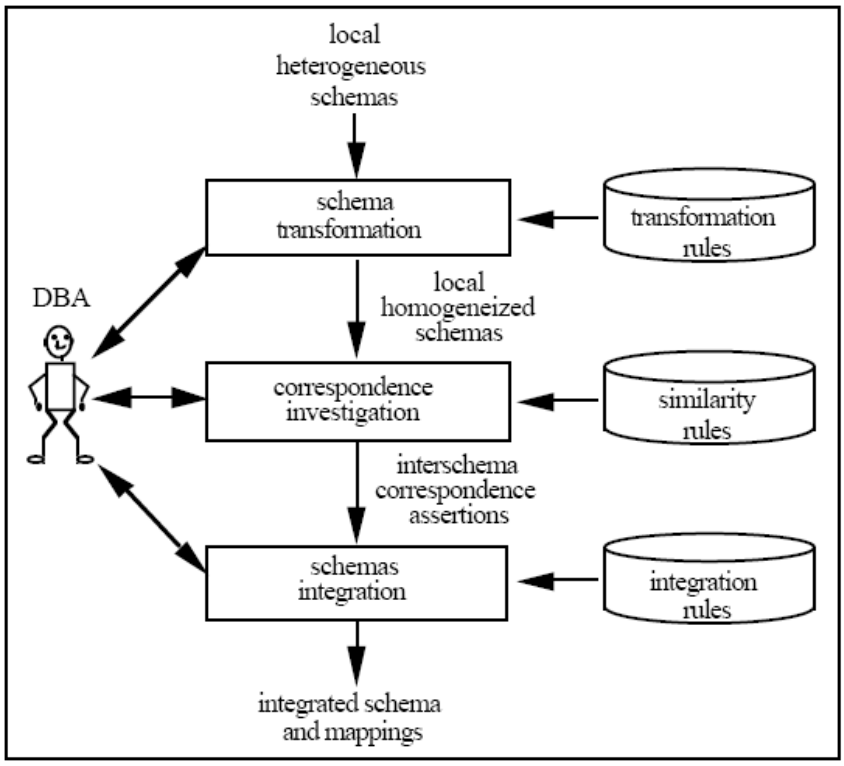


Figure 9 : Processus d'intégration globale [Christine Parent et Stefano Spaccapietra]

Selon nous, l'aide à la validation des données doit intervenir à deux niveaux dans ce processus (figure 9). Tout d'abord, une validation des données dans leur schéma d'origine est nécessaire et ce afin de ne pas intégrer des données erronées. Enfin, une seconde validation doit intervenir au moment où les données sont intégrées dans le schéma final. Dans le cas de l'intégration, on aura uniquement recourt à l'aspect curatif de l'aide à la validation des données. L'objectif principal étant généralement de garantir la continuité des activités par une intégration de données transparentes pour l'utilisateur.

3.3 La migration

Les besoins des entreprises en termes d'outils informatiques plus performants est constant. Qu'il s'agisse de mettre en place de nouvelles fonctionnalités pour les utilisateurs ou de moderniser les outils existants afin par exemple d'accroître les performances, les départements informatiques se

doivent de répondre au jour le jour aux demandes d'un monde en constant changement. Ceci entraîne dans certains cas un remplacement complet des outils informatiques existants. On parle alors de migration.

Une des questions centrales dans ce type d'opération est la reprise des données présentes dans les anciens systèmes. Il s'agit non seulement d'une question centrale mais certainement de l'un des plus gros risques d'une telle opération. En effet, toute altération des données préexistantes peut dans bien des cas engendrer des disfonctionnements au niveau des départements métiers voire dans le pire des cas une interruption de leurs activités. Les raisons peuvent être diverses : perte de données, inconsistances… . Une migration est un *challenge* considérable pour une entreprise tant les volumes de données sont souvent gigantesques et qu'ils proviennent d'environnements divers.

L'importance des risques liés à une migration et leurs coûts font que les entreprises rechignent à procéder de la sorte en préférant reporter de telles opérations au moyen de multiples corrections ou adaptations des systèmes existants communément appelés *request for maintenance*.

Nous distinguons ici deux types de migration ayant toutes les deux un impact non négligeable au niveau des données.

Le premier type est un changement au niveau des systèmes supportant la couche applicative comme par exemple le système de stockage des données (passage d'un système de stockage de données de type fichier vers un système de gestion de bases de données relationnelles) ou encore le changement du système d'exploitation. Quelles sont les raisons de procéder à de tels changements? D'un point de vue technique, Davis Loshin [David Loshin, 2001] présente les raisons suivantes :

1. « Le système de gestion de bases de données est obsolète ;
2. Il n'y a plus assez de personnel pour supporter le système actuel ;
3. La société, qui a mis en place le système a fait faillite ou a abandonné cette activité ;
4. Il y a de meilleurs systèmes de gestion de bases de données disponibles ;
5. Le système de gestion de bases de données n'est pas supporté par les nouveaux ordinateurs ;

6. Il y a de meilleures façons de gérer les données dans un nouveau système de gestion de bases de données. »

Le principal critère de réussite de ce type de migration est de ne pas altérer les fonctionnalités existantes c'est-à-dire de maintenir le même niveau de service qu'auparavant. Dans ce cas la migration est sous la responsabilité unique du département informatique et ne doit avoir d'implications que pour lui. On se situe donc à un niveau plus technique mais nécessitant néanmoins une validation des données avant la mise en production du système. Cette validation n'a pas pour objectif d'améliorer la qualité des données mais de vérifier que les données migrées dans la nouvelle structure n'ont pas été altérées et permettront d'assurer la continuité des activités de l'entreprise. On parle alors de migration iso-fonctionnelle, transparente pour l'utilisateur.

Le second type consiste au remplacement d'un ou de plusieurs systèmes applicatifs par un nouveau système intégré proposant dans la majorité des cas des fonctionnalités additionnelles ou améliorées. Nous abordons dans ce cas les migrations applicatives. Celles-ci ne sont pas transparentes pour l'utilisateur, elles sont principalement réalisées à sa demande et sont sous sa responsabilité. Dans ce cadre, la nécessité d'une haute qualité des données est généralement de mise et est de plus en plus souvent mentionnée dans le cahier des charges. L'aide à la validation des données joue donc ici un rôle central. Notons que ce type de migration peut être combiné avec une première phase qui consiste à intégrer les données.

David Loshin [David Loshin, 2001] décrit les migrations de données comme une recherche archéologique: « nous devons comprendre l'historique de l'utilisation, de la manipulation et du stockage de l'information dans un système de données et ensuite utiliser cette connaissance afin de faciliter le passage aux nouveaux systèmes ». En effet, dans le cas de migration d'anciens systèmes, les standards qui doivent être respectés ne sont parfois plus connus. Des procédés de retro-ingénierie seront alors utilisés afin de retrouver les contraintes auxquelles étaient soumises les données dans l'ancien système. Les contraintes ainsi récupérées devront néanmoins être analysées afin de ne pas répéter les erreurs du passé ou encore répliquer des contraintes obsolètes.

Notons enfin que dans le cas d'une migration, les entreprises doivent voir un tel projet comme une opportunité exceptionnelle de faire un grand pas en avant dans le domaine de la validation des données. Une migration

donne l'opportunité de profiter de l'expérience et des erreurs rencontrées dans le passé afin de mettre en place une validation préventive des données.

3.4 *Reporting* et *data warehousing*

Sur base de ses données, une entreprise génère de nombreux rapports destinés à évaluer de manière continue l'état de l'entreprise (le niveau des services fournis, l'état des ventes…). Ces rapports sont également dans certains cas destinés à répondre aux demandes du législateur. Qu'il s'agisse du premier type de rapport ou du second, la nécessité de fournir des données valides est cruciale. En effet, ces nombreux rapports sont à la base des décisions stratégiques qui permettront à l'entreprise de croître. En ce qui concerne les rapports légaux, si la qualité de ceux-ci est médiocre, l'entreprise peut faire l'objet d'amendes ou d'actions en justice.

Les entreprises ont donc bien compris l'importance de pouvoir avoir à tout moment une vision claire de leurs activités et ont largement investi dans des *data warehouse* littéralement *entrepôts de données* pour répondre à ce besoin. Ils permettent de fournir une vision *cross*-applications de leurs métiers. Ces outils chargent et rafraîchissent de manière continue un nombre considérable de données provenant de sources multiples. La probabilité que certaines de ces sources contiennent des données erronées est donc élevée [Erhard Rahm et Hong Hai Do]. Comme nous l'avons dit, la qualité des données présente dans de tels systèmes est primordiale. L'aide à la validation joue donc un rôle central dans de tels outils.

Un des processus centraux dans l'alimentation d'un *data warehouse* est le processus ETL à savoir *Extraction, Transformation and Load*. Il représente donc les trois étapes essentielles et communément admises pour alimenter un *data warehouse*.

- *Extraction*

« *L'extraction des données est la première étape dans les systèmes ETL. Elle permet de lire les données à partir des systèmes sources. Selon la nature de ces systèmes sources (24/7, critique...) l'extraction peut s'avérer critique et très exigeante dans le sens ou il faut la réaliser le plus rapidement possible et ce en exploitant au minimum les ressources du système source. En général, les extractions sont lancées la nuit durant ce que l'on appelle une extract window. La complexité de l'extraction n'est pas dans le processus de lecture, mais*

surtout dans le respect de l'extract window. D'une part, c'est pour cette raison que l'on effectue rarement des transformations lors de l'extraction. D'autre part, on essaye au maximum d'extraire seulement les données utiles (les données mises à jour ou ajoutées après la dernière extraction). Par ailleurs, pour ne pas perdre des données suite à des problèmes d'extraction, il est important de s'assurer que le système source ne purge pas les données avant que l'entrepôt ne les ait extraites. » [SystemeETL.com, 2009]

- *Transformation*

La transformation est certainement la tâche la plus complexe. En effet, c'est lors de cette étape que l'on intègre les données. On doit donc agréger des données qui à la base appartenaient à des structures différentes.

Cette intégration oblige une analyse minutieuse de la qualité des données qui sont traitées telle que l'analyse des possibles doublons comme par exemple un client présent dans deux applications différentes. Nous pouvons également être confrontés à des données à priori semblables mais qui ont des significations différentes (le pays « Belgique » est représenté par la valeur « 1 » dans un système et la valeur « 2 » dans un autre).

C'est également à ce stade qu'on intègre les règles définies par les départements métiers qui permettront de générer les rapports souhaités (le total des ventes est la somme des ventes depuis le 1er janvier de l'année courante du magasin A et du magasin B). Dans certains cas et pour des raisons de performance, seul le résultat de la règle sera chargé dans le *data warehouse* (dans notre exemple le total des ventes).
Nous citons ici quelques-unes des grandes fonctionnalités de transformation :

- « *Filtrage des données*
- *Nettoyage des données*
- *Standardisation des données*
- *Merging*
- *Mise en conformité des données* » [SystemeETL.com, 2009]

- *Load*

La dernière phase du processus consiste à charger les données dans le *data warehouse*. En fonction des configurations, les données chargées viendront s'ajouter ou écraseront les données préexistantes.

Notons que le processus ETL peut également être utilisé dans d'autres circonstances comme lors de projets d'intégration de données. Dans ce cas, la destination finale des données ne sera probablement pas un *data warehouse*.

La figure 10 montre le processus continu d'alimentation d'un *data warehouse.*

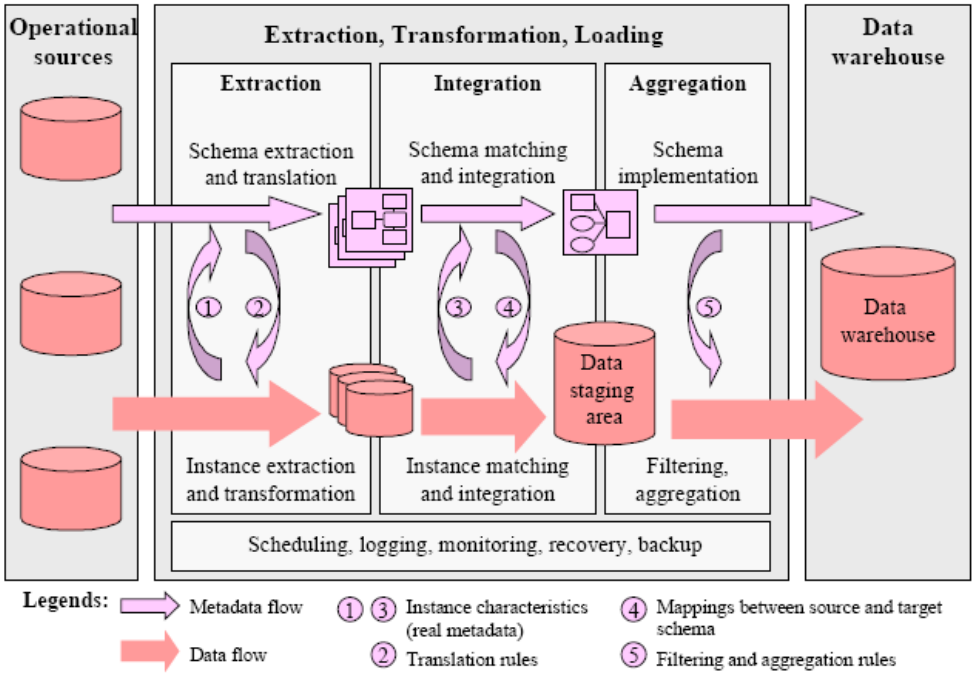


Figure 10: Data warehouse et le processus ETL [Erhard Rahm et Hong Hai Do]

Comme mentionné, l'aide à la validation des données interviendra dans la phase Transformation du processus ETL. Celle-ci sera appliquée dans une Staging Area. C'est-à-dire, non pas sur les bases de données sources ni même sur le data warehouse mais dans une zone « neutre » permettant de limiter les risques d'altération irrémédiable des données et d'éviter les problèmes de performance en particulier sur les systèmes de production.

L'aide à la validation des données dans le cadre des *data warehouse* consistera à appliquer des méthodes curatives principalement automatisées. Il s'agira dans la majorité des cas de programme *batch*.

Enfin, mentionnons que meilleure sera la qualité des données sources, plus le processus ETL sera facile à réaliser.

3.5 Synthèse

Les cas d'utilisation présentés précédemment nous montrent qu'en définitive l'aide à la validation des données doit être envisagée dans tous projets informatiques. Néanmoins, pour des raisons de coût ou de temps, cette problématique n'est parfois pas abordée. Pourtant, le fait de mettre cet aspect de côté peut engendrer des coûts ou des pertes de temps bien plus importants à posteriori.

Nous avons également constaté au travers de ces cas que l'approche de la validation des données peut prendre un aspect plus curatif ou plus préventif. La forme et les techniques de validations varieront donc en fonction des cas d'utilisation.

Nous devons également mettre en évidence deux problèmes récurrents qui sont des obstacles à la mise en place d'outils techniques d'aide à la validation des données et qui apparaissent à plusieurs reprises dans les cas d'utilisation présentés: la performance et la multiplicité des sources des données.

- La performance

Les problèmes de performance sont un point d'attention particulier et engendrent dans certains cas des comportements risqués telle que la désactivation des contraintes référentielles des bases de données. Même si les technologies actuelles sont en constant développement et si des systèmes de plus en plus rapides sont proposés, ceux-ci restent très coûteux et donc souvent laissés de côté.

- La multiplicité des sources de données

L'aspect multi-source des problèmes de qualité des données est particulièrement important dans le sujet qui nous préoccupe. « Selon

IDC, une société, en moyenne, n'a pas moins de 49 applications tournant sur 14 bases de données clientèles différentes » [traduit de Rich, 2008]. Notre expérience nous a montré que l'aspect multi-source et la diversité de ces sources (SGBD, systèmes de fichiers,…) pour lesquelles aucune documentation des modèles physiques ou logiques n'existe ou, si la documentation existe, ont des modèles de données très différents, sont couramment rencontrés dans les sociétés. Cet aspect est certainement un des problèmes et un des défis majeurs pour les migrations, les intégrations de données ou encore la mise en place d'un *data warehouse*.

La diversité des sources de données accroît la complexité de la validation des données. Ainsi, les auteurs E. Rahm et Hong Hai Do dans : « Data Cleaning : Problems and current Approaches » mettent en evidence cet aspect et proposent les distinctions suivantes :

Source unique :

Le premier point abordé par les auteurs est celui d'une seule source de données, c'est-à-dire que toutes les données disponibles au sein d'une organisation sont stockées en un seul endroit.

Ils précisent également que cette source de données unique peut être de différents types. Si de plus en plus de sociétés ont recours au système de gestion de bases de données relationnelles, il existe encore de nombreux environnements utilisant uniquement des systèmes de fichiers voire même n'utilisant que des outils bureautiques tel que *Excel*.

Sources multiples :

Les données dans bien des cas ne proviennent pas d'une source unique mais bien de sources parfois très hétéroclites. Nous avons eu l'occasion de participer à un projet où pas moins d'une dizaine de sources reprenant des données clients avaient été identifiées. De plus, celles-ci étaient de types fort différents allant du fichier Excel aux bases de données relationnelles. L'aspect multi-source aggrave les problèmes rencontrés dans le cas d'une source de données unique.

Enfin, la mise en œuvre de nouveaux projets doit toujours être perçue comme l'opportunité de lancer une révision complète de la gestion des données et de leur qualité ainsi que de sensibiliser le *management* à cette problématique.

2ère Partie : Erreurs et techniques de résolution

4. Les types d'erreurs

Dans ce chapitre, nous présentons de manière très concrète quelles sont les erreurs qui peuvent apparaître au niveau de la donnée. Ceci nous permettra de comprendre de manière pragmatique quels sont les défis qui doivent être relevés dans le cadre de l'aide à la validation des données.

4.1 Exemples

Afin de fixer les idées, nous présentons dans un premier temps quelques exemples de données erronées. Ceux-ci donnerons une première impression de la diversité des erreurs qui peuvent être rencontrées. Pour ce faire, prenons l'exemple simple d'une base de données unique contenant deux tables. L'une représentant les clients, l'autre leurs commandes.

CLIENT

NCLI	NOM	ADRESSE	LOCALITE	CP
1	Goffin C.	72, Rue de la gare	Namur	5000
2	Monti D.	23, Rue Dumont Bruxelles	Bruxelles	1000
3	Toussaint R.	12, Rue Henri	Bruxelles	1030
6	Toussaint R./	12, Rue Henri	Bruxelles	1030
4	Neuman K. (pas sympa)	40, Rue Brasssart	Hotton	6990
4	Hansene G.	112, Rue du printemps, Marche		6900
5	Hansenne G.	112, Rue du printemps	Marche	6900

COMMANDE

NCOM	NCLI	DATECOM
30145	1	10/20/2006
30158	2	
30234	8	24/04/2006
30178	5	12/08/1835

Figure 11: Exemple d'une base de données clients

Sur base de ces seules informations, examinons quelles erreurs peuvent être facilement identifiées :

a. L'attribut NCLI, considéré comme identifiant, présente un doublon.

b. L'enregistrement *Hansene G.* semble être fort similaire à l'enregistrement *Hansenne G.*.

c. *1030* n'est pas le code postal de Bruxelles mais bien celui de Schaerbeek.

d. Une commande a été effectuée en *1835* mais la société n'existe que depuis 10 ans. Cette donnée est donc *out of range*.

e. Nous avons six clients pour quatre commandes. Or pour être client, une commande doit nécessairement être passée.

f. La date de la commande 30158 n'est pas connue.

g. La commande *30234* ne correspond à aucun client connu.

h. L'ajout de la mention *(pas sympa)* est probablement un signal montrant que la structure de données telle que prévue à l'origine ne permet pas l'enregistrement de toutes les données souhaitées par l'utilisateur ou que la structure de l'écran n'est pas claire.

i. La présence d'un enregistrement (table client, id « 6 ») avec un caractère « / » et à priori similaire à un autre enregistrement doit être analysée avec beaucoup d'attention. Il peut s'agir bien entendu d'une faute de frappe, mais ce type d'erreur peut également nous signaler qu'un utilisateur a souhaité de manière délibérée créer un doublon pour une raison quelconque.

j. La localité du client *Hansene G.* ne se trouve pas dans le bon champ.

Les erreurs peuvent ainsi être nombreuses et variées. Nous remarquons également que sur base d'une analyse sommaire de ces erreurs, plusieurs pistes de solutions peuvent être envisagées.

4.2 Typologie des erreurs

Il est important de pouvoir se référer à une classification des erreurs afin de pouvoir les identifier de manière plus aisée et de structurer les actions à entreprendre. Pour procéder à cette classification, nous nous basons sur la typologie proposée par Kim et Coll. dans « Taxonomy for dirty data ». Cette taxonomie est représentée par la figure suivante :

1. Missing data
   - 1.1 Missing data where there is no Null-not-allowed constraint
   - 1.2 Missing data where Null-not-allowed constraint should be enforced
2. Not-missing, but
   - 2.1 Wrong data, due to
     - 2.1.1 Non-enforcement of automatically enforceable integrity constraints
       - 2.1.1.1 Integrity constraints supported in relational database systems today
         - 2.1.1.1.1 User-specificable constraints
           - 2.1.1.1.1.1 Use of wrong data type (violating data type constraint, including value range)
           - 2.1.1.1.1.2 Dangling data (violating referential integrity)
           - 2.1.1.1.1.3 Duplicated data (violating non-null uniqueness constraint)
           - 2.1.1.1.1.4 Mutually inconsistent data (action not triggered upon a condition taking place)
         - 2.1.1.1.2 Integrity guaranteed through transaction management
           - 2.1.1.1.2.1 Lost update (due to lack of concurrency control)
           - 2.1.1.1.2.2 Dirty read (due to lack of concurrency control)
           - 2.1.1.1.2.3 Unrepeatable read (due to lack of concurrency control)
           - 2.1.1.1.2.4 Lost transaction (due to lack of proper crash recovery)
       - 2.1.1.2 Integrity constraints not supported in relational database systems today
         - 2.1.1.2.1 Wrong categorical data (e.g., wrong abstraction level, out of category range data)
         - 2.1.1.2.2 Outdated temporal data (violating temporal valid time constraint; e.g., a person's age or salary not having been updated)
         - 2.1.1.2.3 Inconsistent spatial data (violating spatial constraint; e.g., incomplete shape)
     - 2.1.2 Non-enforceability of integrity constraints
       - 2.1.2.1 Data entry error involving a single table/file
         - 2.1.2.1.1 Data entry error involving a single field
           - 2.1.2.1.1.1 Erroneous entry (e.g., age mistyped as 26 instead of 25)
           - 2.1.2.1.1.2 Misspelling (e.g., principle instead of principal, effect instead of affect)
           - 2.1.2.1.1.3 Extraneous data (e.g., name and title, instead of just the name)
         - 2.1.2.1.2 Data entry error involving multiple fields
           - 2.1.2.1.2.1 Entry into wrong fields (e.g., address in the name field)
           - 2.1.2.1.2.2 Wrong derived-field data (due to error in functions for computing data in a derived field)
       - 2.1.2.2 Inconsistency across multiple tables/files (e.g., the number of Employees in the Employee table and the number of employees in the department table do not match)
   - 2.2 Not wrong, but unusable data
     - 2.2.1 Different data for the same entity across multiple databases (e.g., different salary data for the same person in two different tables or two different databases)
     - 2.2.2 Ambiguous data, due to
       - 2.2.2.1 Use of abbreviation (Dr. for doctor or drive)
       - 2.2.2.2 Incomplete context (homonyms; and Miami, of Ohio or Florida)
     - 2.2.3 Non-standard conforming data, due to
       - 2.2.3.1 Different representations of non-compound data
         - 2.2.3.1.1 Algorithmic transformation is not possible
           - 2.2.3.1.1.1 Abbreviation (ste for suite, hwy for highway)

2.2.3.1.1.2 Alias/nick name (e.g., Mopac, Loop 1, and Highway 1; Bill Clinton, President Clinton, William Jefferson Clinton)
2.2.3.1.2 Algorithmic transformation is possible
2.2.3.1.2.1 Encoding formats (ASCII, EBCDIC, ...)
2.2.3.1.2.2 Representations (including negative number, currency, date, time, precision, fraction)
2.2.3.1.2.3 Measurement units (including date, time, currency, distance, weight, area, volume, ...)
2.2.3.2 Different representations of compound data
2.2.3.2.1 Concatenated data
2.2.3.2.1.1 Abbreviated version (e.g. John Kennedy for John Fitzgerald Kennedy)
2.2.3.2.1.2 Uses of special characters (space, no space, dash, parenthesis, in a social security number or phone number)
2.2.3.2.1.3 Different orderings (John Kennedy vs. Kennedy, John)
2.2.3.2.2 Hierarchical data (e.g. address concept hierarchy: state-county-city vs. state-city)
2.2.3.2.2.1 Abbreviated version
2.2.3.2.2.2 Uses of special characters
2.2.3.2.2.3 Different orderings (city-state, state-city)

Figure 12: Taxinomie des erreurs [Kim et Coll. 2003]

La distinction de base de cette structure est d'une part les données manquantes et d'autre part les données non manquantes mais erronées ou inutilisables. Ils mettent ainsi en évidence certaines erreurs souvent oubliées par d'autres classifications telles que :

• Données étrangères :

Dans ce type d'erreurs, nous avons comme valeur d'attribut plus d'informations que celles désirées. Par exemple pour un attribut qui reprend le nom du client, nous allons retrouver à côté de celui-ci son prénom ou encore sa ville de naissance.

• Données au contexte incomplet:

Prenons par exemple l'attribut « Ville » d'une base de données, on peut y retrouver la ville de Miami, mais s'agit-il de Miami en Floride ou Miami en Ohio?

• Données ordonnées de manières différentes :

On retrouve souvent ce problème dans les champs « Adresse ». En effet, nous pourrions avoir une valeur telle que "Belgique, rue du Luxembourg, 2" et pour un autre enregistrement "rue du Luxembourg, 5, Belgique".

En particulier pour ce type d'erreurs, le contexte d'utilisation joue un rôle majeur. De fait, l'exemple ci-dessus ne posera probablement aucun problème s'il s'agit uniquement d'imprimer l'adresse sur une enveloppe. Par contre, si un programme en amont doit récolter le pays d'habitation des clients afin de fournir des statistiques sur leurs origines en analysant toujours la dernière partie de l'adresse, il sera probablement incapable de le faire avec des données telles que celles-ci. Cet exemple montre toute l'importance de toujours proposer à l'utilisateur des champs aussi structurés que possible.

Notre choix s'est porté sur cette taxinomie pour sa clarté, sa simplicité et surtout la diversité des erreurs qu'elle identifie. En effet, les erreurs sont classées de manière systématique et hiérarchique.
Néanmoins, si la construction de cette taxinomie et son niveau de détails sont particulièrement intéressants, nous en regrettons trois aspects :

- La classification des doublons

Le problème des doublons, qui est certainement l'aspect le plus étudié par la littérature, est ici présenté de manière simplifiée comme étant le résultat d'une faille au niveau des contraintes d'intégrité d'une base de données. Si les bases de données actuelles rejettent facilement deux enregistrements ayant une clé primaire identique, il n'en est pas de même dans les cas bien plus complexes comme présenté à la figure 13.

Il s'agit d'un problème à l'heure actuelle. Nous approfondirons ce sujet au chapitre suivant.

Record	EmpNo	Name	Address
1	142625M	Liu Hang Xiang	1020 Jalan Bandar Lamma, Industrial Park 3, West Malaysia
2	142725M	Mr. Liu H.X.	Ind Park 3, 1020 Jalan Bandar Lama, Malaysia

Figure 13: Exemple de doublons complexes [Mong Li Lee, Hongjun Lu, Tok Wang Ling et Yee Teng Ko]

- La classification des données manquantes

Les auteurs abordent le problème des valeurs manquantes dans le cas où un attribut d'une entité n'est pas complété. Quid alors du cas où c'est un enregistrement complet d'une entité qui est manquant ? Dans ce cas spécifique, il est très difficile de mettre en place des contrôles informatiques efficaces. Ceci ne doit pas empêcher la classification de ces erreurs car des méthodes organisationnelles peuvent limiter ce problème.

- Les données non nécessaires

Le dernier aspect est celui des données qui ne sont pas manquantes ni erronées mais totalement inutiles pour l'applicatif ou l'utilisateur. Notre définition de l'aide à la validation se veut globale et nous sommes donc convaincu que cet aspect doit être également considéré. En effet, il y a un réel impact sur la qualité des données parce qu'il nuit à la performance, qu'il peut amener de la confusion dans l'utilisation des données ou encore qu'il en accroît les coûts de stockage.

Les trois points décrits ci-dessus comme manquant à la classification proposées par les auteurs sont symptomatiques de la littérature actuellement disponible sur le sujet de la qualité des données. En effet, celle-ci est principalement ciblée sur les aspects informatiques du problème. La classification présentée ici n'échappe pas à la règle. Même si les auteurs présentent l'intervention de *Domain Expert* comme inévitable dans la correction des erreurs, la classification proposée se concentre sur les erreurs qui peuvent être traitées ou partiellement traitées informatiquement. Une fois encore les outils informatiques ne sont pas l'unique réponse aux problèmes de qualité de données.

4.3 La profondeur des erreurs

La classification présentée ignore la profondeur des erreurs. Or cet aspect est particulièrement intéressant afin d'identifier l'ampleur des problèmes auxquels nous devons faire face. Nous désignons par le terme profondeur la difficulté à localiser et à corriger un problème de qualité. De manière intuitive, nous pouvons considérer toutes les erreurs relatives au schéma comme des problèmes faciles à résoudre et qui sont d'ailleurs, pour la plupart, souvent vérifiés automatiquement par les SGBD actuels tel que

Oracle 10. Il faut néanmoins être plus prudent avec les problèmes de qualité relatifs aux instances. Leur localisation et leur correction sont beaucoup plus délicates voire impossibles dans certains cas. Nous pouvons donc distinguer au niveau des instances deux types de problèmes :

- Problèmes relatifs à la syntaxe :

Lorsque nous parlons de syntaxe, nous abordons le problème du respect ou non-respect de la grammaire formelle d'un langage [Communauté wikipédia, 22/05/2009]. La syntaxe traite de la forme. Nous parlons donc ici de règles précises qui doivent êtres respectées par une donnée comme par exemple le format d'un compte bancaire. Celles-ci sont dans la majorité des cas facilement interprétables par des programmes informatiques.

- Problèmes relatifs à la sémantique :

La sémantique aborde la signification des données; elle traite du fond. Les problèmes relatifs à la sémantique sont les plus complexes à localiser, à corriger, mais également les plus dangereux. En effet, lors de l'exploitation de ces données sémantiquement fausses peu ou aucun signe d'erreur n'apparaîtra. Par exemple, un enregistrement tel que nom = Mathieu et prénom = Jacques, s'il y a inversion du nom et du prénom, cette erreur sera pratiquement indétectable. Il existe néanmoins quelques solutions aux problèmes de sémantique qui seront présentées dans les chapitres suivants.

5. Les techniques informatiques

5.1 *Data Profiling* et *metadata*: la cartographie des données

Nous définissons le *data profiling* comme une cartographie des données c'est-à-dire une documentation complète et détaillée des données manipulées par les processus métiers. Le *data profiling* fait un usage intensif de métadonnées. Nous entendons par ce terme toutes données décrivant une autre donnée. La documentation que nous présentons dans cette section est donc considérée comme étant un ensemble de métadonnées décrivant les données présentes au sein d'une entreprise. Si au travers de cette documentation nous décrivons la structure, le type et l'utilisation des données en notre possession, nous verrons que les métadonnées peuvent également être utilisées au niveau des instances d'une entité. Nous parlerons alors de *data tagging* (5.4 Le *data tagging*).

Il faut savoir qu'une telle documentation n'existe pas ou seulement partiellement dans le plupart des grandes entreprises. Ceci est notamment dû, d'une part à l'histoire de l'architecture informatique de ces entreprises (fusion, nombre d'années d'existence…) mais également au recours de plus en plus fréquent pour ce type d'entreprise à des **progiciels** ne nécessitant pas, lors de leur mise en place, de documentation détaillée des données gérées puisque celles-ci sont généralement imposées. Or c'est souvent dans ces systèmes, où on a peu de visibilité sur les données, que des problèmes de qualité risquent d'apparaître. En effet, les schémas de données de ce type d'applicatif sont souvent conçus pour garantir un certain niveau de flexibilité au client plutôt que de garantir la qualité des données.

Nous déclinons cette documentation en six documents reprenant des informations techniques mais également des informations métiers. Cette documentation permettra de faire le lien entre les données, les contraintes rencontrées ou imposées par les départements métiers et l'applicatif.

5.1.1 Un schéma conceptuel

Un schéma conceptuel (figure 14) est une représentation des besoins de l'utilisateur en termes de données. Le processus d'élaboration d'un schéma

conceptuel est crucial dans la mesure où il conditionne la qualité d'une base de données [J-L Hainaut, 2002]. Nous pouvons être amené à créer ce schéma après la mise en production de l'application.

Nous insistons sur le fait qu'un schéma conceptuel est une représentation *cross*-applications. En effet, il est courant dans les grandes entreprises qu'une même entité, un client par exemple, soit gérée par différentes applications et/ou bases de données. Afin de pouvoir procéder à la validation de ce type de données, il est important d'appliquer les mêmes règles à toutes ces applications et/ou bases de données afin d'éviter les désynchronisations. C'est notamment pour cette raison que les règles et standards de validation seront décrits au niveau du schéma conceptuel.

Ce schéma nous permettra donc d'avoir un vue globale et compréhensible par les départements métiers des entités (et des attributs) et des relations qu'elles entretiennent entre elles. Sur base de ces informations, nous serons capables d'identifier clairement les données qui devront être validées ainsi que les règles et standards qui devront être pris en compte lors de la mise en place des outils d'aide à la validation des données.

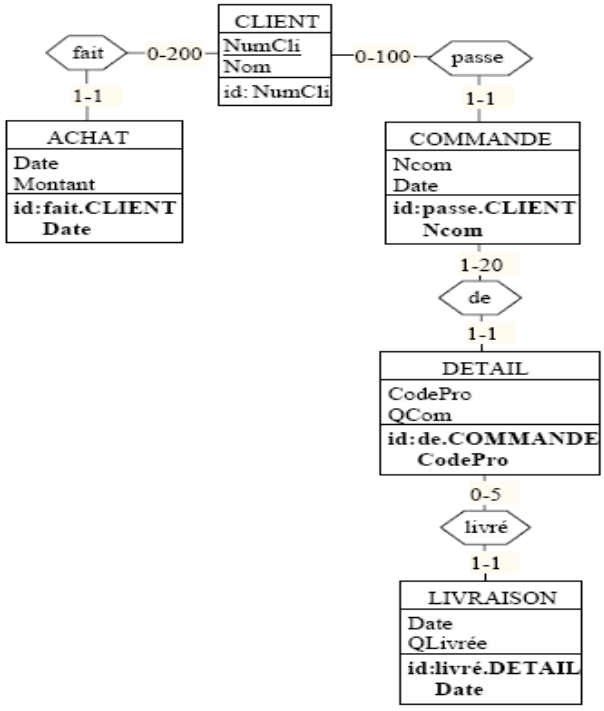


Figure 14 : Exemple de schéma conceptuel [J-L. Hainaut, 2003]

5.1.2 Un schéma physique

Afin de définir ce qu'est un schéma physique (figure 15), nous nous rapportons à la définition proposée par Jean-Luc Hainaut dans le syllabus d'ingénierie des bases de données (2003) :

> « *C'est le rôle du processus de Conception physique que de spécifier les structures techniques et de fixer les paramètres de la base de données et du SGBD de manière à assurer un fonctionnement satisfaisant. Le résultat est appelé le schéma physique.*
>
> *En particulier, on définira les index nécessaires, les espaces de stockage (les fichiers dans lesquels seront stockés les enregistrements) et le mode d'organisation ou de rangement des enregistrements. On définira aussi les techniques d'implémentation des index (B-tree, hashing, inverted lists, etc), leurs paramètres (comme la taille des pages et les taux de remplissage), ainsi que diverses grandeurs définissant le mode de fonctionnement de la base de données et de ses applications (telle que la taille des tampons ou le type des journaux).* »

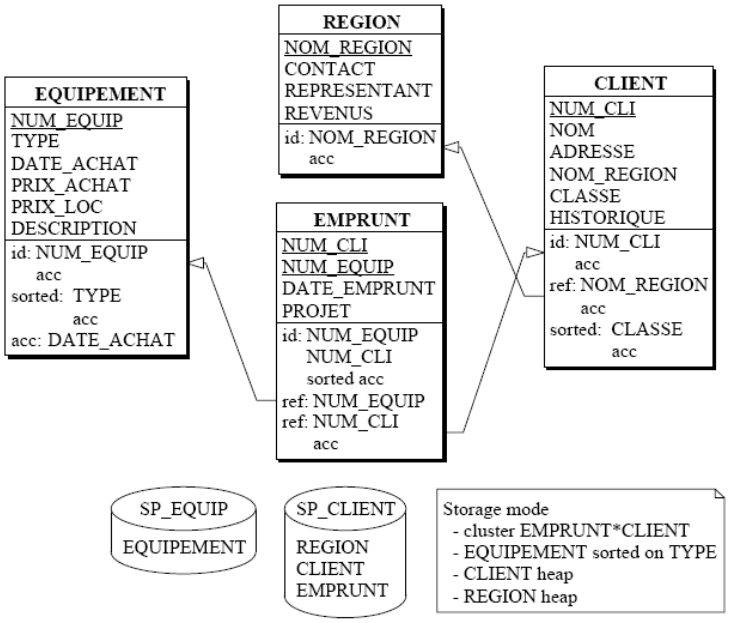


Figure 15 : Exemple de schéma physique [J-L. Hainaut, 2003]

Le schéma physique est bien entendu obligatoire pour mettre en place les outils techniques d'aide à la validation des données, mais également pour vérifier que les contraintes d'intégrité sont bien implémentées. C'est en définitive sur les tables de ce schéma que seront construits les outils de validation.

Enfin, il va de soit que le schéma physique est propre à chaque application contrairement au schéma conceptuel.

5.1.3 Un livre des attributs

Le livre des attributs est un document qui vient en complément au schéma conceptuel et qui permet de faire le lien entre celui-ci et le schéma physique. De plus, ce catalogue a pour but de décrire plus en avant, pour chaque attribut, l'utilisation qui en est faite et quels règles et standards devront lui être appliqués. Bien entendu, l'importance ou la nécessité d'utiliser ces informations complémentaires sera définie par chaque entreprise. Ces informations sont :

- Nom de l'attribut : celui-ci doit être le plus explicite possible et unique pour une entité.

- Obligatoire: nous mentionnons ici si cet attribut doit être obligatoirement présent ou pas.

- Type de donnée : la donnée est-elle numérique, alpha numérique, booléenne… ?

- Longueur de champs : il s'agit de la longueur de l'attribut en termes de nombre de caractères.

- Valeurs possibles : il peut s'agir d'un intervalle de grandeur ou encore d'une liste finie de valeur. On peut également dans le cas de valeur numérique indiquer une valeur minimale ou maximale.

- Définition : nous indiquons ici une description précise de ce que représente l'attribut ainsi que des règles métiers qui y sont associées

- Exemple : nous proposons ici un exemple explicite pour cet attribut. L'exemple permettra souvent de clarifier ce que représente l'attribut.

- Niveau de criticité : le niveau de criticité d'un attribut permet d'évaluer dans quel mesure cet attribut nécessite une attention particulière en termes de validation de données. Nous reviendrons sur ce point dans la définition des objectifs au chapitre 6 (6.3.6).

- Contrôle de qualité : nous mentionnons ici si des contrôles de qualité ont été mis en place pour cet attribut et de quel type ils sont.

- Langage : dans le cas de données alphanumériques, nous mentionnons ici le ou les langues qui seront utilisées pour les valeurs de l'attribut. La langue est souvent un facteur de complexité supplémentaire dans la mise en place des outils de validation.

- Point d'attention : nous indiquons ici les risques sur la qualité qui ont été identifiés pour l'attribut.

- Application « maître » : dans le cas où une donnée est utilisée dans différents systèmes, il est important d'indiquer quel système est « maître » pour cette donnée. Le système « maître » est dans la majorité des cas le système qui crée la donnée.

- Fréquence des mises à jour: on indique ici la fréquence des mises à jour pour l'attribut.

- Destination de la donnée: si la donnée est échangée avec d'autres systèmes, nous indiquons quels sont ceux-ci afin de pouvoir garantir la traçabilité de la donnée.

- Rapport utilisant cet attribut : ce point reprend la liste des rapports utilisant cet attribut permettant ainsi d'identifier les impacts conséquents à une modification de l'attribut

- Représentation de cet attribut dans le schéma physique : nous indiquons ici le lien avec le stockage physique de l'attribut.

5.1.4 *Data provenance* et *data flows*

Si le livre des attributs que nous proposons identifie l'application « maître » d'une donnée, il ne nous permet pas d'avoir une visibilité complète sur le chemin suivi par une donnée au sein d'une architecture informatique. Afin d'identifier comme il se doit ce chemin, nous présentons ici le concept de *data provenance* et la technique des diagrammes de flux de données. Le but de cette documentation étant de **pouvoir, en cas de problème de qualité sur une donnée spécifique, identifier les diverses manipulations** dont elle a été l'objet et donc, de localiser de manière plus aisée l'origine du problème de qualité. En particulier dans les grandes entreprises où l'architecture informatique est particulièrement complexe, une telle description des données est cruciale pour en assurer le suivi et la qualité.

Le concept de *data provenance* est défini par Peter Buneman, Sanjeev Khanna et Wang-Chiew Tan comme référant au « processus de suivi et d'enregistrement de l'origine des données et de leurs mouvements entre les bases de données » [traduit de Buneman et coll.].

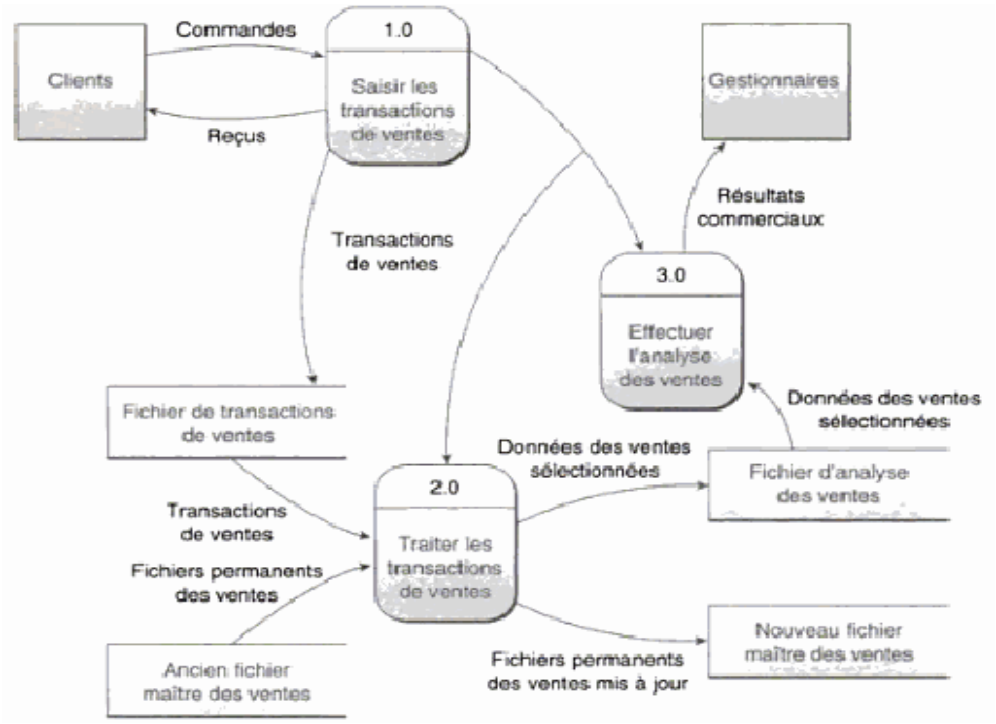


Figure 16 : Exemple de diagrammes de flux de données d'un système de traitement de ventes [Guy Marion, 1997]

Les diagrammes de flux de données peuvent prendre bien des formes que nous ne commenterons pas ici mais l'objectif reste le même. « Un diagramme de flux de données est une représentation graphique des flux de données au travers d'un système d'information » [traduit de Communauté Wikipédia, 20/07/09]. Notons que l'on reconnaît habituellement différents niveaux de détails aux diagrammes de flux de données. La figure 16 nous montre un exemple d'un diagramme de niveau 1.

5.1.5 La matrice CRUD

Le dernier document que nous proposons est la matrice CRUD (figure 17). L'acronyme CRUD désigne les quatre opérations de base sur les données. Soit : Create, Read, Update et Delete. C'est-à-dire : créer, lire, mettre à jour et supprimer. Cette matrice consiste à présenter quels processus métiers agissent, que ce soit en création, lecture, mise à jour ou suppression, sur quelles données.

Cette matrice est une aide très précieuse dans le cadre du sujet qui nous préoccupe. Elle permettra d'identifier les processus et les personnes impactés par les problèmes de qualité de données mais également de mieux comprendre le ou les contextes d'utilisation de cette donnée. Notons que la matrice CRUD reste au niveau des entités pour des raisons évidentes de maintenance.

		Customer	Product	Order	Contract	Part	Supplier	Employee	Financials
Finance	Fin Func 1	C			C,R				C
	Fin Func 2	R		U	C				C
	Fin Func 3	U			U			U	C
Sales	Sales Func 1	R		C		R			
	Sales Func 2				U			U	
Supply Chain	SC Func 1		C,U			R,U	R		R
	SC Func 2		R	R			C		
	SC Func 3						U		R
Etc.									

Figure 17 : Un exemple de matrice CRUD [Jonathan G. Geiger, Intelligent Solutions, Inc.]

5.1.6 Le livre des règles

Une grande entreprise se doit de fixer des règles précises concernant certaines de ses données afin que quelles que soit les applications ou les processus les utilisent, elles restent cohérentes au sein de l'organisation. Par exemple: Est-ce que le numéro de TVA doit être complété pour chaque client ? Quel est le format de date utilisé par l'entreprise ? Quelles sont les conventions de nommage pour une donnée spécifique ? Autant de questions qui devront trouver réponse au sein de ce que nous dénommons un livre de règles qui décrit les standards en matière de données applicables à l'ensemble de l'entreprise et de ses systèmes applicatifs. Sur base de ces règles, l'entreprise envisagera ou non de mettre en place des techniques de validations de celles-ci.

5.1.7 Synthèse

La documentation que nous avons présentée ici est en soi déjà une technique informatique d'aide à la validation des données. En effet, elle permet de déterminer les données sur lesquelles ont doit porter notre attention et qui nécessitent une validation ainsi que de fournir une vision *cross*-applications des données permettant de mettre en place des règles et techniques de validation standardisées au niveau de l'entreprise. De plus, cette documentation fait le lien entre les processus, les applications et les données ce qui fournit une vision globale sur le contexte d'utilisation d'une donnée et les éléments de celui-ci qui pourraient impacter sa qualité. Notons que la documentation présentée, et en particulier le livre de règles, n'est pas figée mais doit au contraire évoluer avec son environnement.

5.2 Le *data cleaning*

Sous les termes *data cleaning* ou *data cleansing* sont regroupés les techniques informatiques permettant de localiser et d'agir sur les erreurs. Ces techniques sont très variées et généralement utilisées dans le contexte curatif de l'aide à la validation des données. Néanmoins, certaines d'entre elles seront la base de techniques préventives.

Les techniques de *data cleaning* sont extrêmement nombreuses car fortement dépendantes du contexte d'utilisation. Nous ne présentons donc ici qu'un petit nombre d'entre elles qui permettront néanmoins au lecteur d'avoir un premier aperçu des analyses que la mise en place de celles-ci

nécessite. Le but n'est pas ici de présenter des techniques de programmation mais bien de décrire le raisonnement qui doit être adopté face à un type d'erreur donné.

5.2.1 La détection des doublons

La détection des doublons est très certainement un des problèmes de qualité de données le plus étudié par la littérature car il s'agit d'un des plus récurrents et des plus coûteux pour les entreprises. Il est souvent une des problématiques centrales dans le cadre d'intégration de bases de données existantes comme par exemple lors de fusions d'entreprises, ou encore dans le cadre d'une migration de plusieurs applications vers une nouvelle application cette fois-ci intégrée. De plus, les techniques utilisées dans la détection des doublons peuvent également être utilisées dans un cadre préventif, le *matching*. Celui-ci consiste à contrôler la validité d'une donnée en la comparant à une ou plusieurs bases de données de référence où l'on recherche un enregistrement identique ou similaire. Cette technique est particulièrement utilisée dans le cadre du contrôle de la validité des adresses. Ces techniques sont également, pour la plupart, aussi utilisées par les moteurs de recherche.

Notons qu'une des principales caractéristiques de localisation des doublons est que ceux-ci ne sont pas abordés au niveau de la valeur d'un attribut mais bien au niveau de l'instance d'une entité.

Nous présentons ici deux méthodes de détection de doublons actuellement disponibles.

5.2.1.1 Détection basique

Les systèmes de bases de données relationnelles actuels permettent de déclarer lors de la création d'une table quels attributs doivent être identifiants (clé primaire) ou uniques (clé secondaire) et ce afin de garantir en partie l'intégrité des données. Néanmoins, ces contraintes établies au niveau de la table ne suffisent pas à éviter l'apparition de doublons. En effet, si ce mécanisme est apte à identifier deux données exactement les mêmes et donc à en rejeter une, il est totalement impuissant face à deux instances similaires d'une entité en termes de données tel que présenté à la figure 18.

4	Hansene G.	112, Avenue du printems	6900 Marche
5	Hansenne G.	112, Rue du printemps	6900 Marche

Clé primaire déclarée au moment de la création de la table dans le SGBD

Figure 18 : Exemple de doublons

Dans ce cas, l'enregistrement sera accepté par le SGBD alors qu'il s'agit d'un doublon. Il convient donc d'appliquer des contrôles supplémentaires. L'approche standard consiste à comparer un enregistrement sur base non pas de un ou deux attributs mais d'une multitude. Chacun de ces attributs sera pondéré afin d'en évaluer son importance dans la détection d'un doublon, comme présenté à la figure 19.

Règle	Pondération
Nom	30
Date de naissance	20
Prénom	15
Lieu de naissance	15
Titre	5
Genre	5
Nationalité	5
Langage	5
TOTAL	100

Figure 19: Exemple de pondération des attributs

Au niveau de l'attribut lui-même, on ne vérifiera pas s'il est identique en tout point à un autre mais on évaluera son degré de similitude, via un certain nombre de règles prédéfinies. Ces règles pourraient par exemple être :

1. Les lettres majuscules ou minuscules sont considérées comme identiques
2. Les différentes déclinaisons d'une lettre ne sont pas considérées comme une différence (e=è=é=ê…)
3. Deux attributs seront considérés comme identiques si 90% de lettres qui les composent sont identiques
4. …

Ainsi, la détection de doublons consiste à confronter un enregistrement à un ensemble d'enregistrements de même type afin de vérifier si l'enregistrement de référence n'est pas un doublon sur base de règles prédéfinies. Notons que l'on peut toujours sur base de règles de priorité stopper la comparaison de deux enregistrements afin d'en améliorer la performance. Il est important de donner des priorités aux règles de comparaison.

5.2.1.2 Détection phonétique

Ce moyen de détection assez récent consiste à se baser sur la phonétique des mots plutôt que sur les caractères qui les composent. Cette méthode permet ainsi de détecter deux mots ayant une orthographe fort différente mais une prononciation similaire comme « orthographe » et « ortograf ».

Cette méthode est particulièrement efficace lorsqu'on aborde la problématique des *call centers*. En effet, les données encodées sont principalement transmises par voix orale ce qui amène de nombreux problèmes de doublons dus à des retranscriptions erronées. Ce mode de détection est évidemment dépendant du langage dans lequel est exprimée la donnée. En Belgique, cet aspect est un frein non négligeable étant donné qu'une majorité des données de nos grandes entreprises sont enregistrées en trois ou quatre langues.

5.2.2 La détection des valeurs manquantes

La détection des valeurs manquantes est en soi assez simple pour autant qu'on sache si l'on doit ou non mettre en place cette détection. Les systèmes de gestion de bases de données actuels offrent des possibilités intéressantes au moment de la création d'une table.

Néanmoins, dire qu'un champ est obligatoire est une chose, s'assurer que son contenu est effectivement différent de *null* (« - », « N/A », « / », …) en est une autre. Prenons un exemple. Supposons que pour pouvoir enregistrer un client dans le système nous devions absolument avoir son numéro de plaque d'immatriculation. Ce champ est obligatoire pour pouvoir lui créer un contrat d'assurance. Si le client ne connaît pas son numéro de plaque mais souhaite contracter le produit, allez-vous réellement lui dire de retéléphoner le lendemain pour faire son contrat ? Non, il sera plus simple de remplir le champ avec une valeur '-' par exemple. De plus, il

arrive couramment que le fait qu'un champ puisse ne pas être rempli est conditionné à la présence d'une valeur dans un autre champ et inversement.

Un programme ayant pour but de détecter ces valeurs manquantes devra donc être capable de répondre dans l'ordre aux questions suivantes :

L'attribut doit-il toujours avoir une valeur ?

1. Si oui et si une valeur est présente, le programme ne retourne pas d'erreur. Dans ce cas, il conviendra de vérifier si la valeur fournie n'est pas assimilée à une valeur *Null* tels que « - », « N/A » ou encore « / ».
2. Si oui et si la valeur est absente, le programme retourne une erreur.
3. Si non, la présence d'une valeur est-elle conditionnée à un événement?
    a. si oui et si l'événement est réalisé, en cas de valeur absente, le programme retourne une erreur ;
    b. si oui et si l'événement n'est pas réalisé, en cas de valeur absente, le programme ne retourne pas d'erreur ;
    c. si non, le programme ne retourne pas d'erreur.

Une fois ces aspects clarifiés, la détection de ces problèmes pourra se faire au moyen de requêtes SQL sur la base de données cible.

Le traitement des valeurs manquantes nous montre une fois de plus que le contexte d'utilisation de la donnée ainsi qu'une description des règles qui y sont attachées sont primordiales pour l'aide à la validation des données.

5.2.3 La détection des erreurs de format

Comme nous l'avons présenté, une donnée syntaxiquement fausse est une donnée qui ne respecte pas une grammaire préalablement définie. Ces données, une fois le format correct déterminé, sont relativement aisées à localiser. La tâche la plus complexe est donc la définition de ce format. Dans ce contexte, les outils informatiques et en particulier les systèmes de gestion de bases de données relationnelles offrent différentes possibilités.

Premièrement, avant de déterminer le format d'une donnée, nous devons préciser son **type** (*boolean*, *integer*, …). Cet aspect est parfaitement géré par les systèmes actuels encore faut il que le contrôle soit activé. De plus, dans de nombreux cas, souvent par facilité ou pour garantir la flexibilité de l'application, le type le plus générique (*text*, *varchar*,…) est choisi, signifiant que tout caractère peut être accepté. Ce sont dans la plupart des

cas ces types mixtes qui posent problème car ils accroissent la complexité de la grammaire. Il faut donc dès la détermination du type d'une donnée prendre en compte l'impact que celui-ci peut avoir sur la mise en place et la complexité des règles de validation.

Deuxièmement, il s'agit de déterminer le **format** d'une donnée et de mettre en place les contrôles associés si ceux-ci sont nécessaires. Pour diriger cette réflexion, nous devons nous poser les deux questions suivantes :

Est-ce que la donnée respecte toujours un format prédéfini ?

1. Si oui, les expressions régulières parfaitement gérées par les langages informatiques permettent alors de vérifier ce format. Notons que certaines valeurs comme un numéro de compte contiennent elles-mêmes une méthode de validation basée sur les expressions régulières.

2. Si non et s'il s'agit en particulier de données de type numérique, il faut se poser la question de l'existence d'un minimum, d'un maximum ou d'un intervalle de valeur et de mettre en place les contrôles qui sont liés à ces éléments.

Est-ce que la donnée correspond toujours à une valeur prédéfinie ?

1. Si oui, il convient de déterminer l'ensemble des valeurs possibles. Si cet ensemble est suffisamment réduit que pour être facilement exploitable par l'utilisateur, on préférera alors l'intégration de celui-ci au niveau de l'interface graphique sous forme de choix proposés. Sinon, on envisagera le contrôle de ces valeurs par programmation. On les comparera alors à l'ensemble de valeurs préalablement défini. Le système de clés étrangères proposé par les SGBD permet aisément de s'assurer que la valeur entrée appartient à l'ensemble des valeurs définies comme correctes.
   Dans le cas des données numériques, on parle plutôt d'un *range* de données valides telles que par exemple, si la donnée est comprise entre 1 et 5, la donnée est considéré comme valide.

Si la réponse aux deux questions précédentes est non, alors aucun contrôle informatique sur le format ne peut être effectué. Notons enfin, qu'aux contrôles mentionnés ci-dessus, il nous faut ajouter un contrôle sur la longueur de champ. Celle-ci peut également être fonction de paramètres.

5.2.4 La détection des erreurs référentielles

Les erreurs référentielles sont un cas typique d'erreurs qui apparaissent dans le cadre d'intégration de bases de données. En effet, les données provenant de bases différentes, les contraintes référentielles entre celles-ci n'existent pas. Dès lors, dans chacune de ces bases, on fait référence dans bien des cas à un même fait du domaine réel. Il s'agit donc d'identifier ces instances qui ont un lien entre elles. Ce phénomène peut également apparaître au sein d'une base de données où aucune contrainte référentielle n'a été établie.

Nous présentons ici une méthode proposée par D. V. Kalashnikov et S. Mehrota [Dmitri V. Kalashnikov et Sharad Mehrotra, 2005] qui consiste à utiliser les relations entre objets pour déterminer la présence de ce type d'erreurs.

Les auteurs partent de l'exemple suivant afin d'expliquer leur approche :

Considérons une base de données (figure 20) reprenant des auteurs et leurs publications. Les auteurs sont caractérisés par les attributs <id, authorName, affiliation>. Les publications quant à elles utilisent les attributs <id, title, authorRef1, authorRef2,…, authorRefN>. Supposons que les enregistrements suivants sont disponibles pour les entités décrites ci-dessus :

$\langle A_1$, 'Dave White', 'Intel'$\rangle$
$\langle A_2$, 'Don White', 'CMU'$\rangle$
$\langle A_3$, 'Susan Grey', 'MIT'$\rangle$
$\langle A_4$, 'John Black', 'MIT'$\rangle$
$\langle A_5$, 'Joe Brown', unknown$\rangle$
$\langle A_6$, 'Liz Pink', unknown$\rangle$

$\langle P_1$, 'Databases ...', 'John Black', 'Don White'$\rangle$
$\langle P_2$, 'Multimedia ...', 'Sue Grey', **'D. White'**$\rangle$
$\langle P_3$, 'Title3 ...', 'Dave White'$\rangle$
$\langle P_4$, 'Title5 ...', 'Don White', 'Joe Brown'$\rangle$
$\langle P_5$, 'Title6 ...', 'Joe Brown', 'Liz Pink'$\rangle$
$\langle P_6$, 'Title7 ...', 'Liz Pink', **'D. White'**$\rangle$

Figure 20 : enregistrements disponibles [Dmitri V. Kalashnikov et Sharad Mehrotra, 2005]

A la figure 20, nous pouvons facilement identifier que l'enregistrement Susan Grey A3, correspond fort probablement à Sue Grey de l'enregistrement P2 grâce à une méthode similaire à celle présentée dans la détection des doublons. Par contre en ce qui concerne D. White (en P2 et P6), il est difficile de savoir si celui-ci correspond à Dave White ou Don White. Afin de lever cette **ambiguïté**, les auteurs vont exploiter les relations existantes.

« Pour ce faire, nous constatons premièrement que l'auteur « Don White » est co-auteur de la publication P1 avec « John Black » qui est du MIT. Nous pouvons utiliser cette information pour lever l'ambiguïté entre les deux auteurs (Dave White ou Don White). En particulier, puisque le co-auteur de « D. White » en P2 est « Suzan Grey» du MIT, il y a une plus grande probabilité que l'auteur « Dave White » en P2 soit « Don White ». La raison de ceci est que les données suggèrent une connexion entre l'auteur « Don White » et le MIT et son absence entre « Dave White » et le MIT

Dans un second temps, nous observons que l'auteur « Don White » a été le co-auteur de la publication P4 avec « Joe Brown » qui est à son tour co-auteur de la publication avec « Liz Pink ». A contrario, l'auteur « Dave White » n'a pas co-écrit de publications ni avec « Liz Pink » ni avec « Joe Brown ». Puisque « Liz Pink » est co-auteur de P6, il y a une plus grande probabilité que « D. White » en P6 réfère à l'auteur « Don White » plutôt qu'à l'auteur « Dave White ». La raison est que dans la plupart des cas les réseaux de co-auteurs forment des groupes/clusters d'auteurs qui réalisent des recherches liées et publient leurs résultats ensemble. Les données suggèrent que « Don White », « Joe Brown » et « Liz Pink » font partie d'un même groupe d'auteurs alors que « Dave White » non.

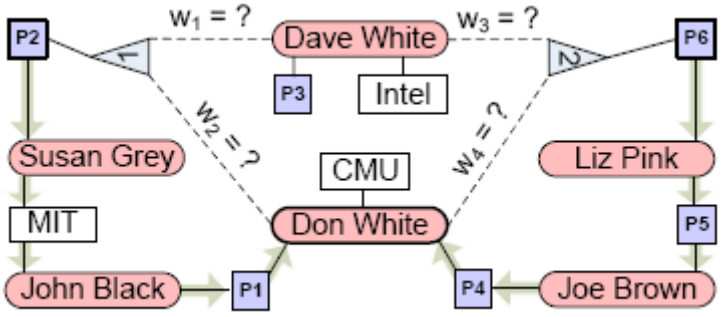


Figure 21 : représentation des relations entre auteurs et publications [Dmitri V. Kalashnikov et Sharad Mehrotra, 2005]

Au premier abord, cette analyse (utilisée pour lever l'ambiguïté sur la référence qui ne pouvait être résolue par une technique de *matching* classique) peut sembler spécifique à un domaine. Néanmoins, un principe général émerge si nous voyons une base de données comme un graphe d'entités interconnectées (les nœuds) liées entre elles par des relations (les arcs). La figure 21 illustre le graphe entités-relations de la base de données prise pour exemple constituée d'auteurs et de publications. Dans ce graphe, les entités qui contiennent des références sont liées aux entités auxquelles elles réfèrent. Par exemple, puisque la référence « Sue Grey » de P2 est résolue de manière non ambiguë comme référent à l'auteur « Susan Grey », la publication P2 est liée par un arc à l'auteur A3. De manière similaire, la publication P5 est liée aux auteurs A5 (« Joe Brown ») et A6 (« Liz Pink »). L'ambiguïté des références « D. White » en P2 et P6 est représentée en liant les publications P2 et P6 à la fois à « Dave White » et « Don White » au moyen de deux nœuds « choix » (1 et 2 dans la figure). Ces nœuds représentent le fait que les références « D. White » peuvent référer aussi bien à l'une ou l'autre entité liée au nœud « choix ».

Etant donné cette vue en forme de graphe, l'analyse qui a été utilisée pour lever l'ambiguïté « D. White » en P2 et P6 peut être vue comme une application du principe général suivant :

> **Principe d'attraction contextuelle :** Si la référence r faite dans le contexte d'une entité x réfère à une entité y_j alors que la description fournie par r correspond à de multiples entités y_1, y_2,… , y_j,…, y_N, alors x et y_j sont probablement plus liés l'un à l'autre via une chaîne de relations que x et y_l. » [traduit de Dmitri V. Kalashnikov et Sharad Mehrotra, 2005]

Sur base de ce principe, nous pouvons mettre en place des algorithmes de parcours de graphe permettant de lever l'ambiguïté de ce type d'erreurs référentielles.

5.2.5 Best Practices : l'ergonomie

L'ergonomie des interfaces utilisateurs a un rôle non négligeable à jouer dans l'amélioration de la qualité des données et donc de ce fait est également une aide à la validation des données. Pour illustrer nos propos, nous prendrons un exemple concret.

Notre premier exemple (figure 22) propose deux façons de présenter une liste de choix de valeurs à l'utilisateur.

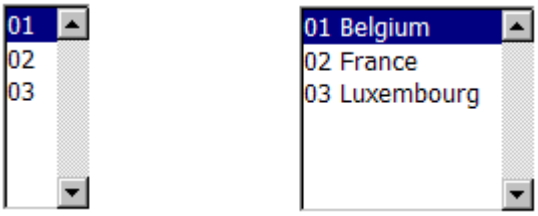


Figure 22: Exemple de listes de choix

Dans la première liste, l'application demande à l'utilisateur de faire référence, de mémoire, à la signification des codes 01, 02, 03 alors que dans le second cas le lien entre le code et sa signification est présenté à l'utilisateur. La première liste risque d'engendrer des erreurs sémantiques difficilement identifiables par la suite.

5.3 Techniques d'action sur les échanges de données

En agissant sur les interfaces, nous ciblons les échanges de données entre applications qu'elles soient internes ou externes à l'entreprise. Nous abordons donc l'aspect multi-sources de l'aide à la validation des données. Comme présenté au chapitre 3, cet aspect accroît considérablement les risques de mauvaise qualité et complique la validation des données. Nous présentons ici les techniques et concepts spécifiques à cette problématique qui sont les plus en vogue aujourd'hui.

5.3.1 Le master data management

Le *master data management* ou MDM est un des concepts les plus discutés aujourd'hui. Afin de bien comprendre celui-ci, il convient de définir ce que l'on entend par *master data*. Pour ce faire, nous nous référons à la définition proposée par Gartner :

> *Master data is the consistent and uniform set of identifiers and extended attributes that describe the core entities of the enterprise and are used across multiple business processes. Some examples of core entities are: parties (customers, prospects, people, citizens, employees, vendors, suppliers or trading partners), places (locations,*

offices, regional alignments or geographies) and things (accounts, assets, policies, products or services).

[Andrew White et coll. 2006]

Nous ajouterons à cette définition que les *master data* sont majoritairement partagées/échangées entre les systèmes informatiques.

Le *master data management* consiste comme présenté à la figure 23 à centraliser la gestion des données utilisées par plusieurs applications afin de pouvoir passer d'un environnement multi-sources à un environnement utilisant une source unique de données qui en sera la référence. Cette centralisation des données facilite la validation de celles-ci en permettant de rationnaliser la mise en place d'outils d'amélioration de la qualité des données ainsi que leur gestion. En particulier, une telle architecture permet une prévention des doublons beaucoup plus efficace grâce à l'élimination par le *master data management* de l'aspect multi-sources de cette problématique. Enfin, le master data management permet dans le cadre d'architectures informatiques complexes la diminution du nombre d'interfaces.

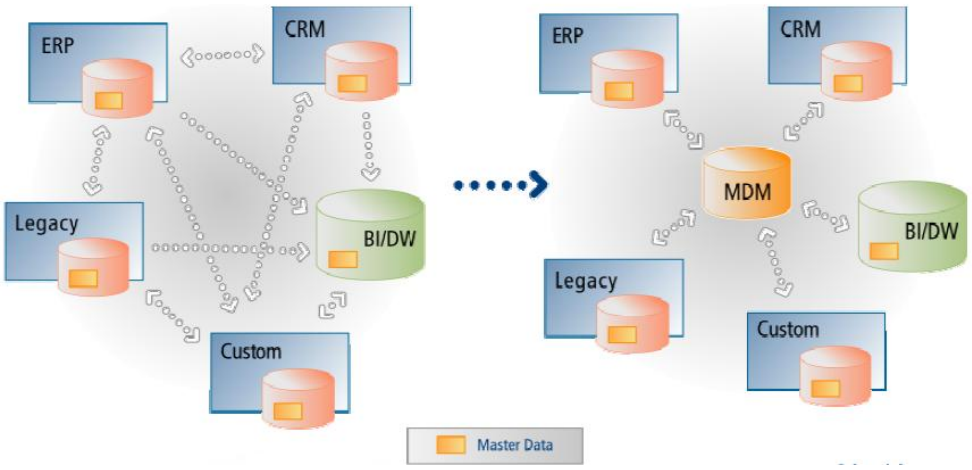


Figure 23 : Master Data Management [Artigas, 2009]

Si ce concept a le vent en poupe, il n'en reste pas moins qu'il présente trois désavantages majeurs :

1. La centralisation des données engendre l'apparition d'un *single point of failure* dans la mesure où elle concentre le risque en un seul point.

2. Les performances en termes de rapidité d'exécution de demandes qui sont adressées à un tel système doivent être très élevées.

3. Le passage d'une architecture décentralisée, souvent très complexe, au concept de data management est très coûteux, risqué et compliqué en termes d'intégration de données qui auront pour la plupart des structures fort différentes à la base.

5.3.2 Le XML

Nous n'entrerons pas ici dans les détails du XML, eXtensive Markup Language, mais nous souhaitons montrer l'importance de celui-ci dans la validation des échanges de données. L'utilisation du XML et en particulier des schémas XML permettent de faciliter les échanges de données ainsi que leur validation. Alors que les documents XLM contiennent les données le schéma XSD qui y est associé « *décrit les éléments qui peuvent apparaître dans ses instances, ainsi que la hiérarchie que ceux-ci doivent respecter. Il précise les éventuels aspects répétitifs ou facultatifs des éléments ainsi que leur type*» [Meurisse, 2004]. Le schéma XSD permet la validation des documents XML au moyen de *parser*. *« Un parser est l'utilitaire de base pour le traitement d'un document XML. Il sert d'intermédiaire entre le document et le programme d'application. Il permet à ce dernier de parcourir le document sur la base de l'arbrorescence formée par les éléments, attributs et données textuelles sans devoir se soucier de la syntaxe XML*. » [Meurisse, 2004].

Il nous semble important d'aborder le XML dans le cadre de ce mémoire tant, s'il est bien utilisé, il permet de garantir un certain niveau de qualité de donnée. Le premier avantage du XML est qu'il est un standard facilitant la mise en place d'autres standards et de ce fait, facilite l'échange de données entre applications, entreprises, etc. Son second avantage est la facilité de sa mise en place grâce aux nombreux *parser* disponibles sur le marché.

Néanmoins, comme nous avons pu le constater dans le cadre du projet SEPA[1], ces avantages peuvent être en partie neutralisés par une mauvaise définition des schémas. Ainsi, dans le cadre de ce projet, le schéma XML

[1] SEPA : Single European Payment Area. Projet européen de standardisation des moyens de paiement. Un des principaux impacts de ce projet pour les banques a été le passage au XML pour les paiements interbancaires.

n'était pas structurant pour l'adresse. Elle était décrite comme deux champs libres de 70 caractères. Il ne fallut pas attendre une semaine après le lancement des virements SEPA pour constater que ces champs étaient extrêmement mal utilisés par de nombreuses banques, le rendant inexploitable par les systèmes applicatifs. Plus un schéma XML sera simple et structurant, plus la qualité des données sera grande.

5.4 Le *data tagging*

Nous présentons ici une méthode permettant de faciliter le processus d'évaluation de la qualité des données. Elle consiste à associer aux instances d'une entité des informations complémentaires permettant d'en évaluer leur qualité. Cette méthode est dénommée *data tagging*.

Ainsi Isabelle Boydens dans Techno 7, n°7 défini le *data tagging* comme suite :

> « *Le data tagging (littéralement « étiquetage des données ») est une méthode développée aux Etats-Unis par le Massachusetts Institute of Technology*6. *La méthode consiste à enrichir le schéma d'une base de données en y ajoutant des informations qui permettent aux utilisateurs d'en évaluer la qualité. L'approche comporte trois étapes :*
>
> - *L'identification des dimensions « subjectives » de la qualité jugées cruciales par les utilisateurs: par exemple, actualité ou encore, fiabilité de l'information ;*
> - *L'identification des indicateurs « objectifs » de la qualité permettant de mesurer certains aspects des dimensions prédéfinies ;*
> - *L'intégration des indicateurs dans le schéma de la base de données.* »

On notera que la mise en place d'un tel outil d'évaluation s'inscrit parfaitement dans l'approche que nous avons de l'aide à la validation des données puisqu'il permet de prendre en compte le contexte d'utilisation.

Ces méthodes sont utilisées de longue date dans l'environnement web. Dans ce domaine, une série de tags standards ont été définis et sont repris

dans le très connu *Dublin Core Metadata Element Set* (DCMES). Cet ensemble est composé de 15 metadata: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

L'utilisation de ces métadonnées doit néanmoins se faire avec parcimonie afin d'éviter tout problème de performance ou même de qualité de ces tags. Pour ce faire, Isabelle Boydens recommande de privilégier comme indicateur de qualité :

- « *des informations déjà présentes dans la base de données : souvent les données s'éclairent mutuellement et leur mise en relation constitue un moyen économique d'en connaître la qualité ;*

- *des données directement générées à partir du système comme, dans notre exemple, les dates d'intégration et de correction de l'information.* »

5.5 Le web, un exemple ?

Le web a certainement été l'un des domaines où la nécessité de la validation des données s'est rapidement avérée comme cruciale. En effet, si dans les applications informatiques classiques on peut dans une certaine mesure compter sur la diligence des utilisateurs, il n'en est pas de même dans le monde de l'Internet. Tout webmaster qui a mis en place un formulaire sans validation des données sur son site internet a connu dès le début de la mise en ligne de celui-ci des déboires tels que des adresses email erronées rendant impossible l'envoi d'email aux visiteurs, des numéros de téléphone ou des adresses totalement fausses. L'internaute rechigne souvent à fournir de telles informations, surtout au début de l'Internet où une grande méfiance régnait vis-à-vis de tels formulaires. Il préférait alors remplir les champs au moyen d'informations ne se référant à rien d'existant. Dans ce contexte, maintenir une base de données cohérente est un vrai défi

On se rend vite compte qu'il s'agit d'une des préoccupations centrales du webmaster. Il suffit de regarder le nombre d'articles consacrés à la qualité des données dans la sphère Internet. De ce fait, les technologies et langages dédiés à l'Internet offrent des possibilités particulièrement intéressantes au niveau de la validation des données. PHP 5 propose ainsi

en standard des méthodes de contrôle des données saisies spécifiques au domaine de l'Internet :

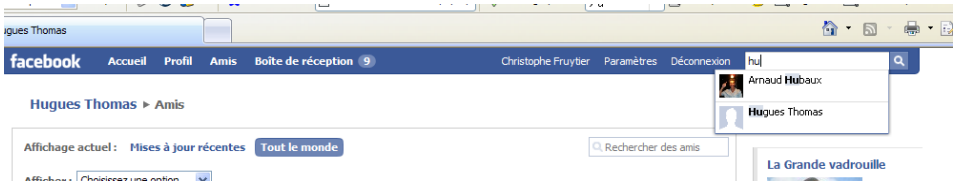
`FILTER_VALIDATE_EMAIL:`
valide une valeur en tant qu'adresse email.
`FILTER_SANITIZE_SPECIAL_CHARS :`
Encode en HTML les caractères ′, ″, <, > et & ainsi que tous les caractères de code ASCII inférieur à 32.

[extrait de Olivier Heurtel, 2007]

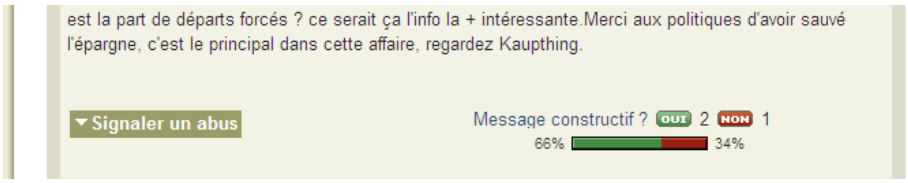
De plus, les gestionnaires de site Internet redoublent d'imagination en ce qui concerne la mise en place de fonctionnalités permettant d'améliorer la qualité des données. Le WEB 2.0 a encore accru les possibilités de validation.

Nous présentons à la figure 24 quelques moyens de validation trouvés sur internet.

www.google.be



www.facebook.com



www.lesoir.be

Figure 24 : Exemples d'aide à la validation des données sur le web

3ère Partie : Mise en place de l'aide à la validation des données

6. Définir

Maintenant que les concepts de base de la validation des données, que les types d'erreurs qui peuvent être rencontrées ainsi qu'un aperçu des techniques informatiques capables de les résoudre ont été décrits, nous allons présenter dans ce chapitre quels sont les points essentiels de la première étape de la mise en place d'outils d'aide à la validation des données. Pour ce faire, ce chapitre sera divisé en trois parties.

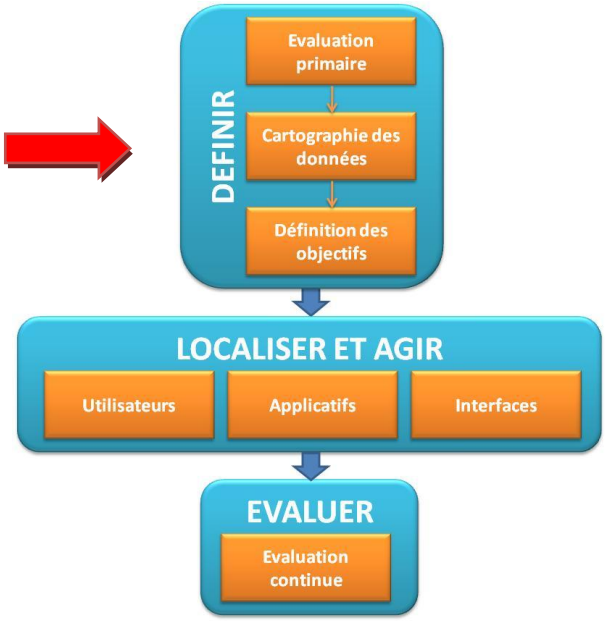


Figure 25: La phase «définir» dans le processus de mise en place d'outils d'aide à la validation des données

Premièrement, nous parlerons de l'**évaluation primaire des données**. Il s'agit de faire un état des lieux afin de déterminer si la mise en place de systèmes d'aide à la validation des données est nécessaire ou non.

Deuxièmement, nous avons proposé au chapitre 5 une **méthodologie de documentation des données**. Cette méthodologie a été discutée sous le point « Data Profiling et metadata : la cartographie des données ». Nous

renvoyons donc le lecteur à ce chapitre concernant ce point. Nous insistons néanmoins sur le fait que cette documentation est un élément primordial dans cette phase de définition. Elle sera la base à la définition des objectifs et dans une phase ultérieure, elle permettra une identification beaucoup plus efficace de l'origine des problèmes de qualité, notamment grâce aux diagrammes de flux de données, et facilitera l'identification des données dont la qualité est à améliorer.

Troisièmement, nous aborderons **la définition des objectifs de qualité** et des standards qui y sont associés.

Dans ce chapitre, nous proposons une approche du problème qui se centre sur les besoins de l'utilisateur final. De manière générale, la définition des standards doit être réalisée en collaboration directe avec l'utilisateur afin tout d'abord de répondre à ses besoins mais également de l'impliquer pleinement dans le processus de qualité des données. De cette façon, l'utilisateur sera moins enclin à détourner le standard défini facilitant ainsi la validation. Les entreprises spécialisées dans le domaine de la qualité des données s'accordent à dire que la phase la plus cruciale dans de tels projets est la définition des objectifs, des procédures, règles et standards qui seront la base de la validation.

Trop souvent dans les projets informatiques, l'accent est mis sur la description des processus tandis que la description des besoins en termes de données arrive dans un second temps. Et même à ce niveau, on se borne à faire une description sommaire des données sans se soucier des problèmes de qualité qui pourraient apparaître. Une description des valeurs valides n'est abordée que lors de la phase d'implémentation.

6.1 L'évaluation primaire

Avant de recourir à des techniques de validation de données, il convient de faire un "état des lieux", c'est-à-dire **réaliser un audit des données en notre possession afin d'en déterminer la qualité**. L'évaluation primaire, comme nous le verrons dans la suite de ce chapitre, sera toujours liée au contexte dans lequel les données seront utilisées.

Cette première analyse est cruciale, car en fonction du résultat, il sera décidé ou non de la mise en place d'une aide à la validation des données. De plus, selon la classe d'erreurs rencontrées, on aura déjà une première orientation du type d'aide à la validation qui devra être mise en place.

Bien entendu, cette première étape ne s'envisage que dans le cas où des données sont préexistantes. Néanmoins, dans le cas particulier de nouvelles applications ne reprenant aucune donnée de systèmes préexistants, nous pensons que les jeux de test devraient être capables de fournir une première mesure de la qualité des données permettant de tirer des conclusions pour le fonctionnement futur du système et si nécessaire l'adaptation des contrôles déjà mis en place.

Pour ce qui est des méthodes d'évaluation, nous prônons tout comme Leo. L. Pipino, Yang W. Lee et Richard Y. Wang dans *Data Quality Assesment* une double approche : d'une part, une **évaluation subjective** et d'autre part une **évaluation objective**. Cette façon de procéder nous permet de prendre en considération le contexte d'utilisation de la donnée et non pas uniquement sa représentation au sein de la base de données.

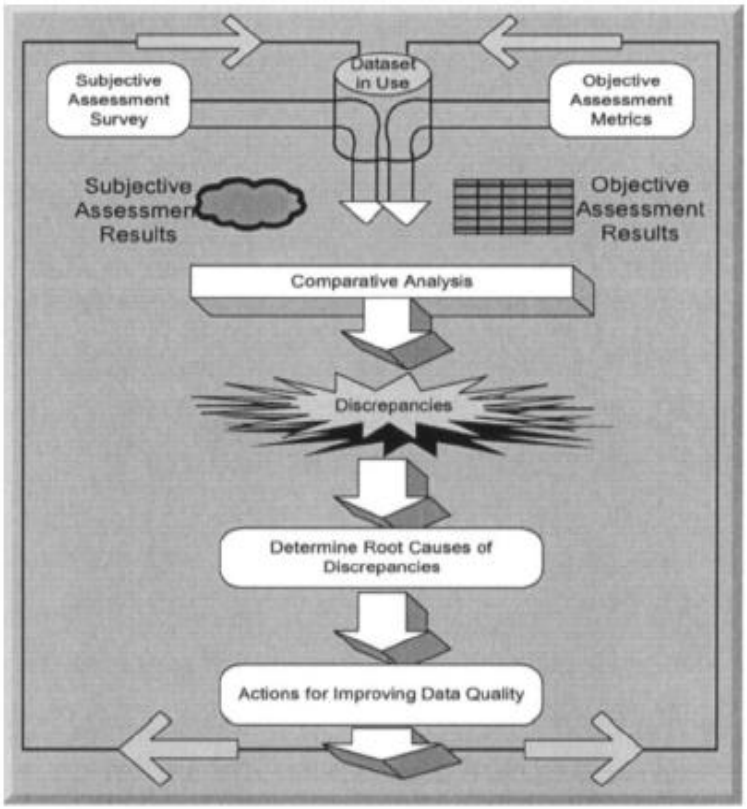


Figure 26: Evaluation primaire [Leo L. Pipino, Yang W. Lee, et Richard Y. Wang, 2002]

6.1.1 L'évaluation subjective

L'utilisateur doit être au centre du processus d'aide à la validation des données. Il est donc logique que l'on se tourne vers lui pour procéder à une première évaluation de la qualité des données et de ses besoins en termes de validation. Nous considérons donc l'évaluation subjective comme une série d'interviews avec l'utilisateur final des données. Cette évaluation doit nous permettre d'atteindre quatre objectifs :

- **Cerner le contexte d'utilisation des données**, en d'autres termes mieux comprendre les processus métiers qui manipulent les données. En effet, dans le cadre des grandes entreprises, les processus, de part leur nombre et dans bien des cas leur complexité, ne sont connus que de ceux qui les pratiquent quotidiennement. Les départements de support, tel que le département informatique, n'ont en définitive qu'une connaissance sommaire des processus métiers et de leurs contraintes.

- **Evaluer la perception des départements métiers des données qu'ils manipulent**.
  Par exemple : Sont-ils attentif à la qualité des données ? Ont-ils des aprioris sur les données qu'ils utilisent ? Doivent-ils régulièrement procéder à des corrections manuelles de ces données ? Sont-elles à jour ? Font-ils régulièrement face à des doublons ?

  Afin de cadrer ces interviews, la présentation de la typologie des erreurs décrites au chapitre 4 ainsi que celle d'exemples concrets peut être un guide.

- **Fournir une première évaluation des coûts engendrés par les problèmes de qualité identifiés**. Une approche assez simple et peu coûteuse de réaliser cet objectif consiste à partir des problèmes générés par une mauvaise qualité des données.

  Prenons l'exemple d'une société spécialisée en *direct marketing*, celle-ci dispose d'une base de données consommateurs reprenant leurs noms, prénoms et adresses. Un bon moyen d'évaluer la qualité de ces données est d'analyser par exemple le nombre de retour courrier. Si 10% des courriers sont refusés par la poste, on peut raisonnablement conclure que la base de données utilisée est

de mauvaise qualité. De nombreux autres exemples tels que le nombre de plaintes adressées à un call center, le nombre de paiements erronés reçus, peuvent être utilisés comme un moyen de mesure.

- **Conscientiser les départements métiers aux problèmes de la qualité des données.** En effet, s'ils doivent faire face à ce problème depuis de nombreuses années, ils ne le considèrent probablement plus comme un problème mais un élément faisant une partie intégrante de leur travail.

Le premier avantage de ce type d'évaluation est qu'elle permet d'avoir une première évaluation de la qualité des données sans développement informatique et donc à moindre coût.

Le deuxième avantage est son impact auprès du management et en particulier le management non-IT. Particulièrement pour le troisième objectif, cette façon de procéder permet aux départements opérationnels d'objectiver dans leurs tâches quotidiennes le coût en temps et en argent de ce que peut représenter une mauvaise qualité de données. Présenter le problème en disant que 10% des données sont erronées est souvent beaucoup trop abstrait et ne permet pas d'obtenir l'implication et l'accord du management pour lancer un programme d'amélioration de la qualité des données. Par contre, l'objectivé en termes de coût pour l'organisation a un impact plus conséquent.

6.1.2 L′évaluation objective

Sur base des informations recueillies lors de l'évaluation subjective, l'évaluation objective consiste à analyser directement les données présentes dans les systèmes. La première étape est d'extraire un sous-ensemble aussi représentatif que possible des données d'un système. Ainsi, si une société est composée de quatre départements de vente (un par segment client), il convient d'extraire de cette base de données clientèles un nombre représentatif de clients par segment, ce nombre devant être proportionnel à un nombre de clients du segment par rapport au nombre total de clients.

Comme il s'agit d'une pré-analyse de la qualité des données et afin d'en limiter les coûts, nous conseillons de procéder à des analyses simples de données: au moyen de requêtes SQL (par exemple identifier les attributs non complétés) ou encore par simple visualisation. La manipulation de ces

données dans un tableur peut également être parfois suffisante pour tirer de premières conclusions.

6.1.3 Interaction entre les deux méthodes

L'évaluation objective doit nécessairement être conduite en supplément de l'analyse subjective. En effet, si l'avis de l'utilisateur est de première importance, il faut être capable de contrebalancer ses dires. De plus, certaines erreurs ne seront peut-être pas identifiées par lui mais pourront l'être lors de l'évaluation objective. Enfin, dans certains cas, une mauvaise qualité de données peut être due à une volonté de l'utilisateur telle que, par exemple, la création volontaire de doublons dans le système pour réaliser des simulations.

De plus, les deux évaluations présentées ici ne doivent pas être réalisées indépendamment. Un aller-retour entre les deux méthodes est nécessaire afin de raffiner progressivement les résultats.

Ces méthodes d'évaluation sont simples et permettent assez rapidement d'avoir une première évaluation de la qualité des données et d'ainsi envisager ou non la mise en place d'outils d'aide à la validation des données.

6.3 La définition des objectifs

Dans le premier chapitre, nous avons donné une définition de la qualité des données. Nous avons précisé que ce concept était multidimensionnel. Nous avons également présenté que dans le cadre de ce mémoire, nous analysions l'aide à la validation des données dans le cadre de l'amélioration de la qualité de celles-ci. Cette section a pour objectif de présenter les principales dimensions proposées par la littérature. Celles-ci, sur base de l'évaluation primaire et de la cartographie des données, permettront de faciliter la définition des objectifs de qualité que doit se fixer l'aide à la validation des données. Elles fourniront un cadre aux discussions. Enfin, nous démontrerons une fois de plus que le contexte d'utilisation a une influence indéniable sur les objectifs de qualité.

6.3.1 Les dimensions dominantes

Les dimensions les plus fréquemment citées par la littérature sont l'exactitude, la fiabilité, l'exhaustivité, la pertinence, et enfin l'intemporalité (figure 27). Parmi ces cinq dimensions nous en retenons quatre : l'exactitude, l'exhaustivité, la pertinence, et enfin l'intemporalité. En effet, nous considérons qu'une donnée exacte est fiable et inversement.

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Figure 27 : Dimensions principalement citées par la littérature [Wand et Wang, 1996]

Nous définissons celles-ci de la manière suivante :

6.3.1.1 Pertinence

Au travers de cette dimension, fréquemment citée, nous nous intéressons au fait que les données stockées doivent correspondre aux besoins de l'utilisateur. C'est-à-dire que celui-ci sur base des données fournies puisse en extraire l'information souhaitée. Notons ici la nuance entre donnée et information : les **données** constituent le matériel brut, désorganisé. « Lorsque les données sont traitées, organisées, structurées ou présentées dans un contexte défini de manière à les rendre directement utilisable, on parle alors d'**information »** [traduit de Diffen, 18/04/2008].

Cette dimension est par définition fortement dépendante du contexte d'utilisation. Ainsi, une base de données démographiques (âge, sexe, date de naissance...) de la population belge sera tout à fait pertinente pour évaluer le vieillissement de la population mais très insuffisante si nous devons calculer l'évolution de l'emploi dans les années à venir.

Notons que la pertinence d'une donnée prédomine les autres dimensions décrites ci-dessous. En effet, si une donnée n'est pas considérée comme pertinente les autres dimensions ne doivent plus être envisagées.

6.3.1.2 Exactitude

Cette dimension est la plus étudiée par la littérature mais également la plus complexe. Il n'y a pas ici de définition communément admise. Nous définissons cette dimension comme suit: une donnée est exacte lorsqu'elle représente l'état d'une entité du domaine réel telle qu'observée, sans altération. La donnée doit être fidèle à l'observation faite dans le monde réel à un instant *t*.

Considérons une base de données qui contient les noms, adresses, numéros de téléphone et adresses email des médecins de l'état du Texas. Nous savons que cette base de données contient un certain nombre d'erreurs, certains enregistrements sont faux. Si nous comparons cette base de données à la véritable population de médecins, nous estimons qu'elle a une exactitude de 85%. Si cette base de données est utilisée par l'état du Texas pour prévenir les médecins d'une nouvelle loi concernant la mort assistée, elle sera certainement considérée comme de qualité médiocre. Il serait dangereux d'y recourir dans ce cas particulier. Si cette base de données est utilisée par un fabriquant d'un nouvel appareil chirurgical afin de trouver des clients potentiels, elle sera considérée comme étant de très bonne qualité. [Jack E. Olson, 2003]

Nous constatons donc qu'un même ensemble de données peut être considéré de faible qualité pour une certaine utilisation et de très bonne qualité pour une autre.

6.3.1.3 Exhaustivité

Au travers de cette dimension, nous voulons nous assurer que l'ensemble des observations du monde réel soit repris. Aucune donnée ne doit être absente.

Considérons une base de données qui contient des informations sur les réparations effectuées sur les équipements d'une entreprise. Il est connu que certaines de ces réparations sont effectuées mais que celles-ci ne sont pas enregistrées dans la base de données. Ceci est le résultat d'un manque de temps de certaines personnes qui effectuent les réparations ainsi que d'un manque de surveillance de la part des superviseurs. Le montant

d'informations manquantes est estimé à 5%. Cette base de données est probablement de bonne qualité pour évaluer la santé globale des équipements. Cependant, essayer d'utiliser cette base de données comme support pour l'outil de facturation amènerait indubitablement au non paiement de certaines réparations effectuées. [Jack E. Olson, 2003]

Une fois encore, le constat est le même. Cette base de données ne présentera pas le même niveau de qualité dans un cas ou dans l'autre.

6.3.1.4 Intemporalité

Cette dimension aborde le problème de la mise à jour des données. Une donnée de qualité se doit de représenter le dernier état connu d'une entité du domaine. Nous nous référons ici à la période de temps nécessaire entre l'apparition d'un nouvel état pour une entité donnée et la mise à jour de cette entité dans le système de gestion de base de données.

Considérons une base de données contenant les ventes d'une division d'une entreprise. Cependant, certaines équipes doivent mettre cette base de données à jour en fin de mois, d'autres quotidiennement ou de manière hebdomadaire. Si on utilise cette base pour calculer les bonus des vendeurs qui sont dus les 15 du mois, cette base sera considérée comme de mauvaise qualité. Par contre, s'il s'agit de l'utiliser comme base pour calculer les tendances du marché et envisager des solutions stratégiques, alors la base de données sera considérée comme de bonne qualité. [Jack E. Olson, 2003]

6.3.4 Autres dimensions proposées par la littérature

Comme déjà décrit, la qualité des données est présentée comme un concept multidimensionnel et plusieurs centaines de dimensions ont déjà été identifiées. Dans le cadre de ce mémoire, nous en avons sélectionné quatre et ce pour trois raisons :

- celles-ci étaient les plus couramment citées par la littérature ;
- selon nous, elles permettaient de guider au mieux la réflexion dans 4 axes que nous considérons comme majeurs ;
- ces axes sont suffisamment génériques pour être appliqués dans la majorité des cas.

Néanmoins, nous présentons à la figure 28 un aperçu de quelques autres dimensions. Celles-ci peuvent être intéressantes dans certains contextes d'utilisation spécifiques.

.

Dimensions	Definitions
Accessibility	the extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	the extent to which the volume of data is appropriate for the task at hand
Believability	the extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which data is compactly represented
Consistent Representation	the extent to which data is presented in the same format
Ease of Manipulation	the extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which data is correct and reliable
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial
Relevancy	the extent to which data is applicable and helpful for the task at hand
Reputation	the extent to which data is highly regarded in terms of its source or content
Security	the extent to which access to data is restricted appropriately to maintain its security
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand
Understandability	the extent to which data is easily comprehended
Value-Added	the extent to which data is beneficial and provides advantages from its use

Figure 28 : Exemple d'autres dimensions [L. L. Pipno,Y,. W. Lee and Y. Wang, 2002]

6.3.5 Corrélation avec le contexte d'utilisation

En analysant, les dimensions proposées ci-dessus, nous nous rendons compte que la qualité des données dépend avant toute chose de l'utilisation qu'on en fait. De mêmes données peuvent avoir un niveau de qualité totalement différent d'une utilisation à une autre. Si tel est le cas quel est l'intérêt d'aborder ce concept au travers de différentes dimensions ? Nous donnons trois raisons à cela.

Premièrement, il s'agit d'objectiver dans la mesure du possible une notion subjective. Ceci implique qu'on ne pourra jamais totalement caractériser le concept de qualité des données. Ce phénomène est mis en évidence par le nombre important de dimensions proposées par la littérature. Deuxièmement, ces dimensions aident à une meilleure compréhension du concept mais également à cibler les problèmes majeurs de qualité qui peuvent apparaître au sein d'une base de données.
Troisièmement, ces dimensions donnent un canevas, des points de repère au processus d'amélioration de la qualité des données.

Cette vision de la qualité des données implique donc logiquement que toute démarche effectuée dans le but de l'améliorer doit être réalisée en étroite collaboration avec l'utilisateur.

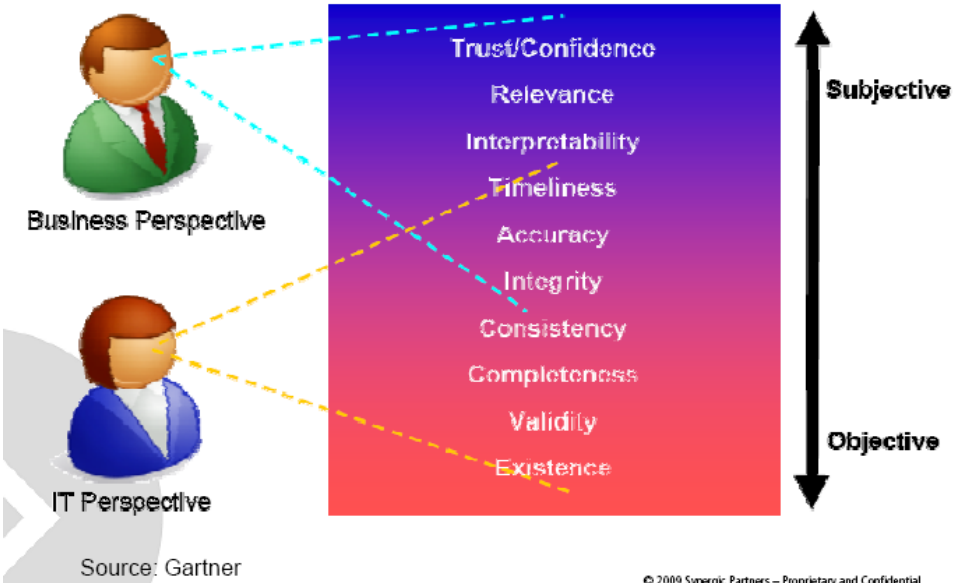


Figure 29: Vision du métier et de l'IT sur les dimensions de la qualité des données [Artigas - Synergic Partners, 2009]

Enfin, nous mettons en évidence la différence de perception de l'importance des dimensions qu'il peut exister entre les départements informatiques et les départements métiers. Cette différence est présentée à la figure 29. Ce schéma confirme la nécessité que la gestion des données soit abordée de manière conjointe par ces départements. Nous présenterons dans le chapitre suivant comment impliquer ces deux mondes dans un projet d'aide à la validation des données.

6.3.6 En théorie

6.3.6.1 Evaluation de la criticité d'un attribut

Nous proposons ici une méthodologie pour définir les objectifs de qualité à la fois au niveau des entités et des attributs. Pour ce faire, nous nous baserons sur les dimensions présentées précédemment ainsi que sur la cartographie des données décrites au chapitre 5.

Tout d'abord, il doit être décidé au niveau de l'entreprise quelles sont les dimensions de la qualité des données qui sont importantes pour elle. Nous conseillons de choisir un nombre restreint de ces dimensions afin de ne pas créer de confusion.

Une dimension sera néanmoins toujours présente : la pertinence. En effet, comme mentionné auparavant, cette dimension impacte toutes les autres. Elle indiquera l'importance d'une donnée pour un domaine d'activité déterminé. Par exemple, si une base de données clientèle récolte le groupe sanguin de chaque client, on conviendra que cette donnée doit être dans tous les cas exacte. Par contre si la collecte de cette donnée n'est pas pertinente et n'est donc pas utilisée dans le contexte d'utilisation qui nous occupe, l'importance de cette donnée sera nulle. Dans ce cas précis, on peut se demander si cet attribut doit être conservé.

Ensuite, pour chaque entité, on pondérera les dimensions choisies. Ainsi, une dimension peut être totalement neutre pour une entité donnée. Sa pondération sera alors de 0.

L'étape suivante consiste à définir pour chaque attribut d'une entité le pourcentage des données qui doivent être en ligne avec la dimension analysée. Par exemple, si on analyse la dimension « intemporalité» et que l'on fixe l'objectif à 80%, cela signifie que 80% des données au moins doivent être à jour. Notons que lorsque, de par la définition de l'attribut,

une dimension est toujours remplie, l'objectif sera fixé à 0%. Pour illustrer ce cas reprenons l'exemple de la date de naissance. Celle-ci ne change jamais. Dés lors, l'objectif de la dimension « intemporalité » sera fixé à 0%. Ce faisant, nous diminuons la criticité de l'attribut. En effet, celui-ci ne nécessitera pas d'intervention sur sa qualité en ce qui concerne cette dimension.

Enfin sur base de ces informations, on peut alors calculer le niveau de criticité de l'attribut au moyen de la formule suivante :

$$C = \sum_{I=1}^{n} (O_n \times E_n) \times P$$

Où :

O représente l'objectif défini pour la dimension n au niveau de l'attribut exprimé en pourcentage

E représente la pondération de la dimension *n* au niveau de l'entité exprimé en pourcentage

P représente la pertinence de l'attribut exprimée en pourcentage.

C représente le niveau de criticité de l'attribut. Plus la valeur de C est proche de 1, plus l'attribut est considéré comme critique et **nécessite donc une attention particulière quant à sa qualité**.

Notre choix s'est porté sur cette formule pour les raisons suivantes :

1. La moyenne pondérée : elle permet d'envisager un nombre variable de dimensions déterminées par l'entreprise (On) et pondéré en fonction de son contexte d'utilisation (En). Le choix de la moyenne pondérée nous permet donc d'adapter l'évaluation d'un attribut à son contexte d'utilisation.

2. La multiplication par P : Comme présentée, la pertinence est une dimension particulière, toujours présente contrairement aux autres et qui peut annuler la criticité d'un attribut. Nous la sortons donc de la moyenne pondérée et nous utilisons la multiplication afin de pouvoir gérer ces deux caractéristiques.

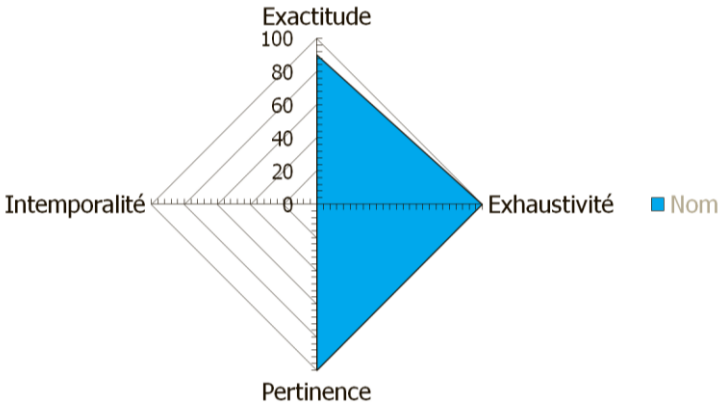
6.3.6.2 Exemple d'utilisation

Considérons une entreprise qui a décidé de mettre l'accent sur les quatre dimensions que nous avons identifiées comme les plus citées par la littérature à savoir la pertinence, l'exactitude, l'exhaustivité, et l'intemporalité. Nous décidons de déterminer les objectifs pour l'entité « Client » du CRM de cette entreprise. Cette entité reprend les attributs suivants : le nom, le genre, l'adresse, le commentaire et le groupe sanguin du client. Pour cette entité la pondération des dimensions est la suivante :

0,4 pour l'exactitude ;
0,3 pour la complétude ;
0,3 pour l'intemporalité ;

Sur base de ces informations, nous avons établi pour chaque attribut son niveau de criticité :

- Le nom



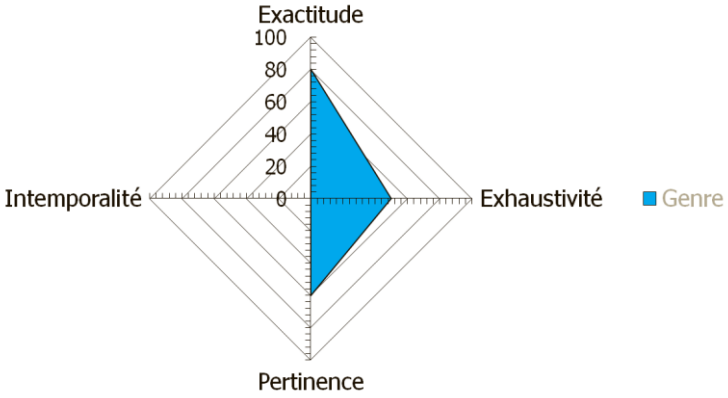
Pour cet attribut, sa pertinence et les objectifs par dimension ont été définis de la manière suivante :

	Nom
Exactitude	90
Exhaustivité	100
Pertinence	100
Intemporalité	0

Son niveau de criticité est donc de :

$$C_a = 0,66 = ((0,9x0,4)+(1x0,3)+(0x0,3)) \times 1$$

- Le genre



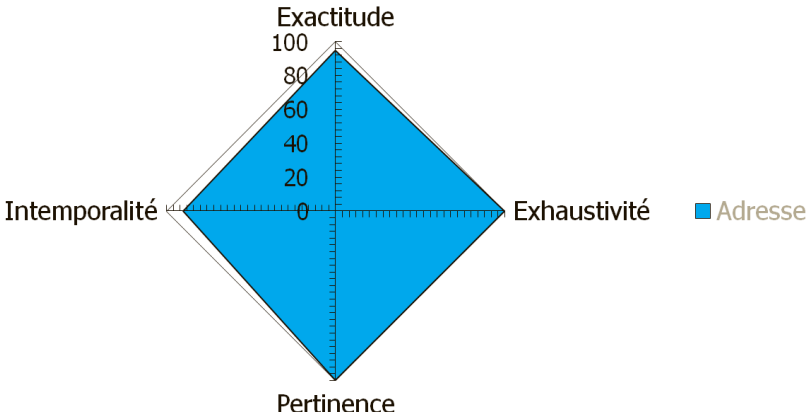
Pour cet attribut, sa pertinence et les objectifs par dimension ont été définis de la manière suivante :

	Genre
Exactitude	80
Exhaustivité	50
Pertinence	60
Intemporalité	0

Son niveau de criticité est donc de :

$$C_a = 0,282 = ((0,8x0,4)+(0,5x0,3)+(0x0,3)) \times 0,6$$

- L'adresse



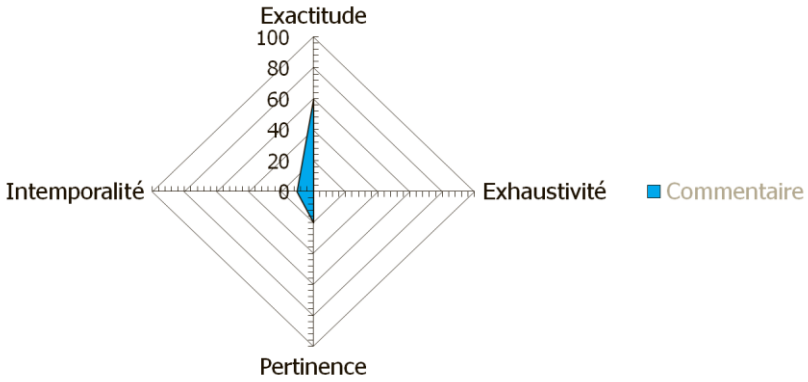
Pour cet attribut, sa pertinence et les objectifs par dimension ont été définis de la manière suivante :

	Adresse
Exactitude	95
Exhaustivité	100
Pertinence	100
Intemporalité	90

Son niveau de criticité est donc de :

$$C_a = 0,95 = ((0,95x0,4)+(1x0,3)+(0,9x0,3)) \times 1$$

- Le commentaire



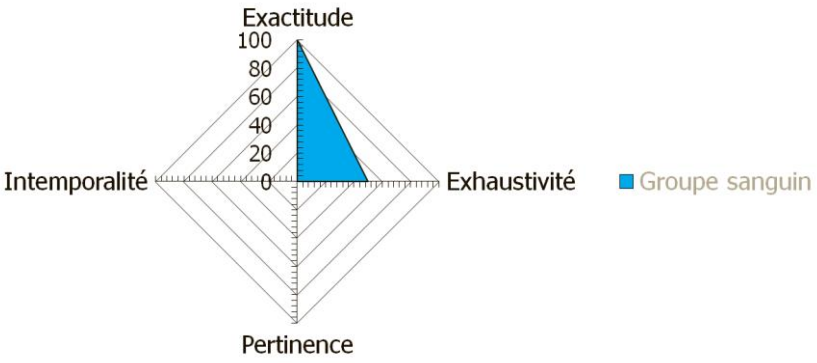
Pour cet attribut, sa pertinence et les objectifs par dimension ont été définis de la manière suivante :

	Commentaire
Exactitude	60
Exhaustivité	0
Pertinence	20
Intemporalité	10

Son niveau de criticité est donc de :

$$C_a = 0,054 = ((0,6x0,4)+(0x0,3)+(0,1x0,3)) \times 0,2$$

- Le groupe sanguin



Pour cet attribut, sa pertinence et les objectifs par dimension ont été définis de la manière suivante :

	Groupe sanguin
Exactitude	100
Exhaustivité	50
Pertinence	0
Intemporalité	0

Son niveau de criticité est donc de :

$$C_a = 0 = ((1x0,4)+(0,5x0,3)+(0x0,3)) \times 0$$

Pour conclure cet exemple, nous établissons le niveau de criticité de l'entité qui est la moyenne des niveaux de criticité des attributs. Dans cette moyenne, les attributs dont la criticité est de 0 ne sont pas repris et ce afin de ne pas masquer la criticité des autres attributs qui composent l'entité. Ainsi, dans ce cas, l'entité a un niveau de criticité de:

$$C_e = 0,4865 = (0,66 + 0,282 + 0,95 + 0,054)/4$$

Dans cet exemple, nous avons opté pour une représentation graphique particulière. Celle-ci a l'avantage d'être suffisamment flexible que pour gérer un nombre indéfini de dimensions. On remarquera que plus la surface délimitée par les objectifs des dimensions (représentée ici en bleu) est grande, plus la criticité de l'attribut est importante. Néanmoins, cette

représentation présente une faiblesse dans le cas où la pertinence est nulle. En effet, elle ne permet pas d'annuler la surface. Cependant, dans ce cas particulier, l'attribut devrait être écarté car il est inutile.

Enfin, on notera, qu'étant donné que le nombre de dimensions n'est pas fixe, il n'est pas possible d'établir une typologie des formes crées par cette représentation graphique.

6.3.6.3 L'apport du modèle

Nous avons précisé au chapitre 1 qu'une qualité parfaite, à de rares exceptions près, ne devait pas être un objectif en soi. Il convient donc de classifier et prioritiser les attributs et entités nécessitant une attention particulière ou une intervention quant à leur qualité. Dans la plupart des cas, cet exercice se réalise de manière intuitive ou encore de manière réactive lorsqu'un problème est rencontré.

Le modèle que nous proposons permet de rationnaliser et de formaliser cet exercice de prioritisation notamment en permettant une comparaison des attributs et des entités sur bases de critères standards à l'entreprise. Ce modèle permet donc d'allouer plus efficacement les budgets destinés à la qualité des données et permet de fixer des objectifs à atteindre.

Nous ne procéderons pas à la validation du modèle proposé dans le cadre de ce mémoire. Nous tenons néanmoins mettre en évidence deux risques que peut comporter l'utilisation de ce modèle et qui devront faire l'objet d'une analyse plus approfondie. Il s'agit d'une part de la charge de travail que peut représenter une telle analyse et d'autre part du risque d'une définition faussée des objectifs par certains départements dans le but d'obtenir des budgets plus conséquents. Dans ce dernier cas, la mise en place d'une cellule de *data management* peut être une solution. Nous reviendrons sur ce type d'organisation au chapitre 9.

7. Localiser et agir

La définition de la validation des données qui a été proposée était l'évaluation d'une donnée par rapport à un standard défini. Nous allons dans ce chapitre voir par quels moyens nous pouvons confronter les données à ces standards, afin de localiser et d'agir sur les erreurs.

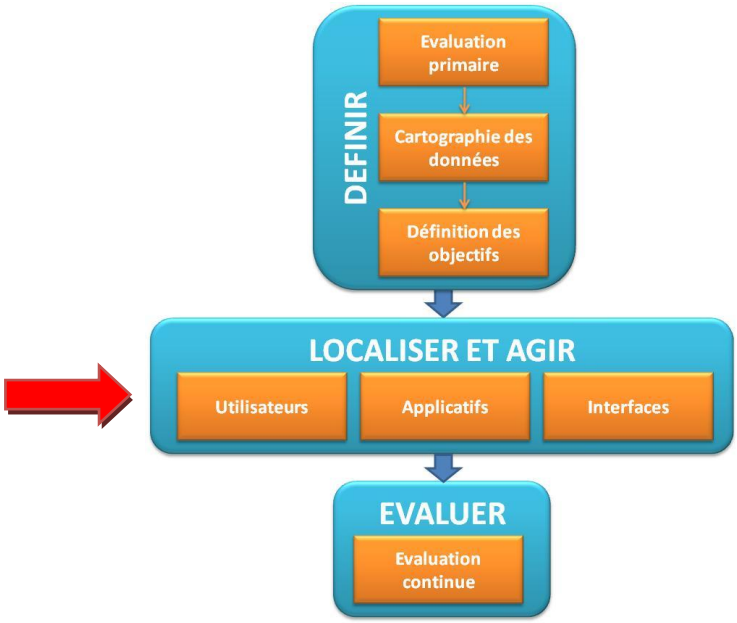


Figure 30: La phase de «localiser et agir» dans le processus de mise en place d'outils d'aide à la validation des données.

7.1 Domaines d'actions

Afin de pouvoir cibler et traiter, de manière curative ou préventive, les différentes erreurs présentées dans le chapitre 4, il est important de pouvoir connaître et comprendre leurs origines.

Dans un premier temps, il faut se rendre à l'évidence que les causes amenant à une faible qualité des données sont très nombreuses. L'esprit humain n'a en effet pas son pareil pour les multiplier. Nous proposons néanmoins une classification permettant de structurer l'approche à mettre en

place. Nous sommes parti du postulat suivant : tous problèmes de qualité des données sont générés par un élément qui a la possibilité de créer, mettre à jour et/ou supprimer ces mêmes données. Nous considérons dans le cadre de ce travail que la lecture de données ne présente aucun risque et que donc cette action est totalement neutre pour la qualité des données.

La littérature identifie généralement deux domaines principaux à l'origine des erreurs, l'application et l'utilisateur. Nous en identifions ici un troisième : les interfaces.

- Applications :

L'aspect applicatif des problèmes de qualité des données est certainement le plus étudié. Certains disfonctionnements, erreurs de programmation ou même l'ergonomie de l'application peuvent générer des données erronées.

- Utilisateur :

L'utilisateur est souvent le premier à se plaindre des problèmes de qualité des données et pourtant il en est souvent la cause. Que ce soit par son manque d'implication dans la définition des fonctionnalités souhaitées ou par le manque d'attention accordé aux données saisies dans son travail journalier (champs non complétés ou incorrectement complétés, création volontaire de doublons…). Lorsque les actions que doit effectuer l'utilisateur final sont répétitives et/ou doivent être exécutées dans un délai restreint, ces problèmes sont accrus.

- Interfaces systèmes :

Des applications totalement isolées sont difficilement envisageables aujourd'hui. Les échanges de données sont donc également une cause de pollution des données si ceux-ci ne sont pas surveillés. S'il est vrai que les erreurs présentes dans ces échanges sont aussi générées par l'homme ou l'applicatif, les entreprises n'ont que très peu de prises sur ces deux éléments lorsque ces interfaces sont établies avec le monde extérieur. Elles doivent donc mettre en œuvre des solutions spécifiques d'où notre volonté d'identifier ce point comme une cause d'erreur distincte. Concernant ce point, nous renvoyons le lecteur au chapitre 5, point « 5.3 Techniques d'action sur les échanges de données ».

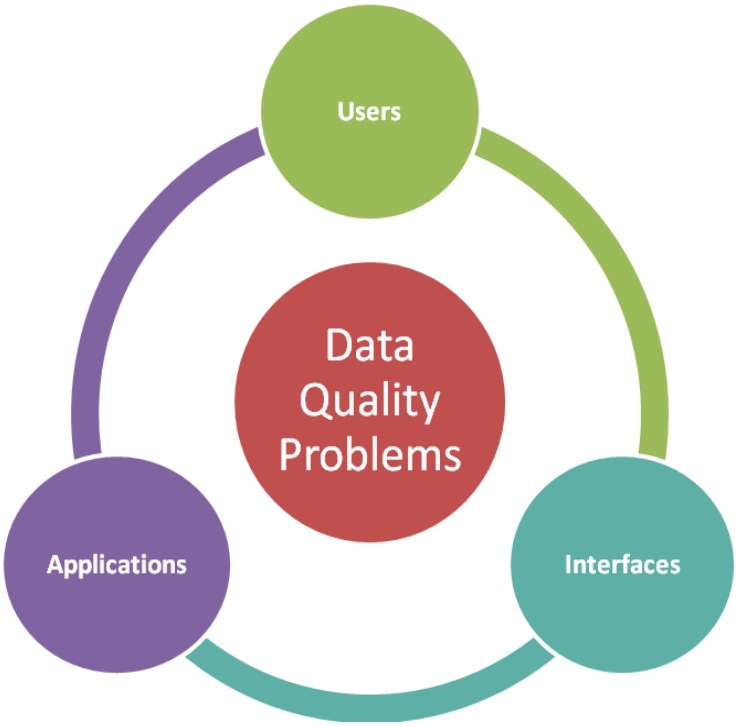


Figure 31 : Domaines d'actions

Ces trois domaines présentés à la figure 31 permettent la classification des origines des erreurs.

Notons que les études actuelles se centrent particulièrement sur les solutions techniques à mettre en place au niveau de la sphère applicative. Nous souhaitons, dans la suite de ce chapitre, mettre également l'accent sur les solutions organisationnelles qui peuvent être mises en place. Nous sommes convaincu qu'une amélioration de la qualité des données au sein d'une organisation doit d'abord passer par une conscientisation et une implication des départements métiers dans la gestion de la validation des données.

En ce qui concerne les solutions applicatives permettant d'améliorer la qualité des données qui, selon nous, doivent venir en support des solutions organisationnelles que nous allons présenter ici, nous renvoyons le lecteur au chapitre 5, point «5.2 Le *data cleaning* ».

7.2 Agir sur l'utilisateur

Dans cette section, nous aborderons les solutions organisationnelles actuellement proposées en termes de validation de données. Ces méthodes sont nombreuses et variées et souvent très spécifiques. Par conséquent, nous ne présenterons qu'un panel restreint des méthodes organisationnelles d'aide à la validation ;

Si les moyens techniques ont un rôle non négligeable à jouer dans la validation des données, ils ont cependant un inconvénient majeur dans le sens où ils déresponsabilisent l'utilisateur, celui-ci ayant la fausse impression que les systèmes informatiques sont là pour prévenir ou corriger toutes ses erreurs. Il est donc important de **remettre la responsabilité de l'utilisateur au centre du processus de validation**. Nous présentons donc ici des méthodes permettant de prévenir les erreurs en agissant directement sur l'utilisateur final d'une application. De ce fait, nous présenterons uniquement le cas précis des méthodes préventives d'aide à la validation des données.

7.2.1 La définition de procédure

Une définition claire de procédures décrivant précisément les étapes d'un processus et les responsabilités de chacun permet de manière générale d'accroître la qualité dans un processus industriel. Lorsque la qualité des données est également abordée dans ces procédures, on a alors un outil terriblement efficace pour prévenir les erreurs.

Une des procédures les plus connues dans ce domaine est certainement celle communément appelée le *four eyes principle* particulièrement répandue dans le monde bancaire et plus spécifiquement dans les moyens de paiement. Elle consiste à faire intervenir deux personnes lors de l'encodage de données. La première a pour rôle d'encoder les données à proprement parler. La seconde, quant à elle, vérifie la qualité des données encodées par la première. Bien entendu, cette procédure est supportée par des règles de sécurité et d'accès fournies par l'applicatif. L'inconvénient de cette procédure est son coût en termes de ressources humaines mais elle permet d'atteindre un niveau de qualité de données très élevé.
Cette façon de procéder a été renforcée dans les grandes entreprises avec la réglementation SOX (*Sarbanes-Oxley Act*) introduite aux Etats-Unis en

2002 qui a fait de la ségrégation des tâches, en particulier au niveau des infrastructures informatiques, un des moyens de contrôle majeur.

7.2.2 La documentation : modèles, règles et standards

Nous nous référerons au chapitre 5 en ce qui concerne la documentation dans le cadre de l'aide à la validation des données.

7.2.3 La sensibilisation et la formation

Comme nous l'avons dit, la validation des données est une problématique qui est présente dans tous les départements d'une entreprise et à tous niveaux. Il est dès lors primordial de sensibiliser un certain nombre de personnes à cet aspect de la gestion des données. Cette sensibilisation permettra de faciliter la mise en place des outils de validation de données mais également conscientisera et responsabilisera les utilisateurs. Ceux-ci n'ont généralement que peu d'informations sur les impacts que peut avoir une mauvaise saisie de certaines valeurs.

Certaines entreprises ont été au-delà de la simple sensibilisation en formant quelques personnes par département afin qu'elles puissent prendre le rôle de *data steward*. Un *data steward* est un utilisateur formé au concept de qualité de données. Celui-ci informera et répondra aux questions de ses collègues sur la qualité des données. Il sera surtout capable de détecter le risque de problèmes de qualité au sein de son département.

Si l'utilisateur final doit être sensibilisé à la problématique de la validation des données, il en est de même pour le personnel des départements informatiques. Ils doivent comprendre les impacts financiers qu'une mauvaise qualité des données peut engendrer. Ils doivent également comprendre l'importance d'intégrer le concept de la qualité des données dans les analyses fonctionnelles et techniques des applications qu'ils seront amenés à créer. Ceci est particulièrement vrai pour les personnes en charge de la modélisation des bases de données.

7.2.4 La communication

Comme nous l'avons déjà dit, investir dans la qualité des données, c'est signer pour un CDI. Il s'agit donc d'informer continuellement les

collaborateurs sur l'importance et l'état de celle-ci. Ceci permet de maintenir l'attention sur cet aspect et de faciliter la transmission d'informations de l'utilisateur aux personnes adéquates.

7.2.5 L'appui du Top Management

Au delà du fait que le management vote le budget pour mettre en place des outils de validation, un message clair de celui-ci sur l'importance d'avoir des données valides a parfois un impact beaucoup plus fort.

De manière plus radicale, un directeur déclarait avoir trouvé une façon infaillible de garantir une très haute qualité de ses données :

On lui annonça un jour que la qualité de ses données clients était catastrophique. Il convoqua ses responsables de vente et leur annonça que si la qualité des données de leurs clients respectifs n'était pas parfaite endéans les deux semaines, il suspendrait leur bonus. Deux semaines plus tard, la qualité des données clients de cette entreprise n'était plus remise en question.

8. L'évaluation continue

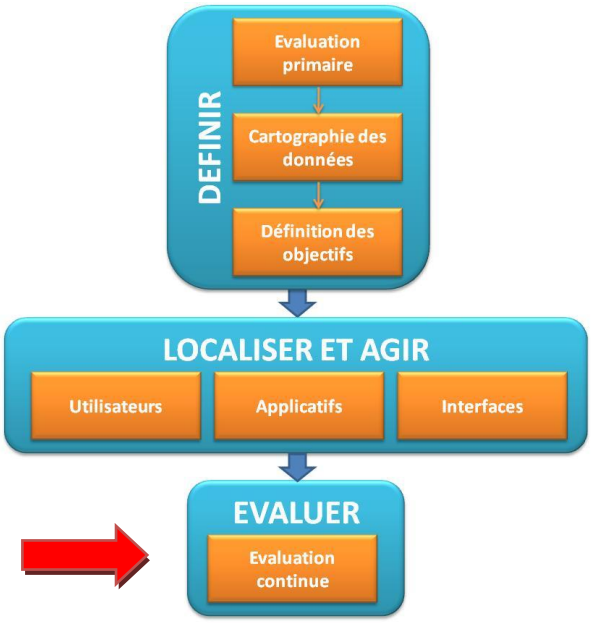


Figure 32: La phase «Evaluer» dans le processus de mise en place d'outils d'aide à la validation des données

« La qualité d'un produit dépend du processus tout au long duquel le produit est façonné. De la même manière, la qualité des données dépend du *design* et du processus de production impliqués dans la création de la donnée. » [traduit de Wand et Wang, 1996]. Afin de détecter les problèmes de qualité et y remédier, ce processus doit être continuellement évalué. Dans ce chapitre, nous présentons l'évaluation de ces problèmes.

L'évaluation continue est, comme son nom l'indique, un processus récurrent et qui est effectuée après la mise en place des outils d'aide à la validation des données. Il permet de contrôler l'état de la qualité des données et d'évaluer continuellement les solutions qui ont été mises en place ou qui devront être mises en place ainsi que le respect des objectifs de qualité qui ont été fixés.

Pour ce faire, l'évaluation continue doit selon nous mettre en place trois types de rapports :

- Les rapports opérationnels

  Ces rapports consistent à renseigner les impacts qu'ont eu la mise en place d'outils d'aide à la validation des données sur les activités métiers d'une entreprise. On n'analysera donc pas ici les données elles-mêmes mais les conséquences de l'amélioration ou de la détérioration de la qualité de celles-ci. On sera par exemple attentif, tout comme lors de l'évaluation primaire, à la diminution, stagnation ou augmentation des plaintes des clients, ou encore aux corrections manuelles de certaines données.

- Les rapports spécifiques

  L'analyse s'effectue ici au niveau de la donnée elle-même. Bien entendu toutes les données ne peuvent être évaluées. Il convient donc de cibler les données à risque ou encore les données qui ont fait l'objet de la mise en place d'outils d'aide à la validation afin de confirmer ou d'infirmer la complétion des objectifs qui avait été fixée pour ces outils.

  Ainsi, par exemple, il est courant de rencontrer des attributs dont la valeur est déterminée par une liste proposée à l'utilisateur. On retrouve souvent la valeur « autre » au sein de ces listes. Le risque d'une telle valeur est qu'elle devienne une valeur « poubelle » où l'on place tous les enregistrements par facilité. Ce type de comportement doit donc pouvoir être contrôlé au moyen de rapports ciblant spécifiquement ce type de valeur à risque.
  De la même manière, il est intéressant de pouvoir contrôler la date de la dernière mise à jour de certaines données. Des données qui n'ont plus été mises à jour depuis plusieurs années alors qu'elles sont susceptibles de changer tous les mois doivent être identifiées et contrôlées.

- Les rapports techniques

  Les défaillances techniques d'une application peuvent avoir dans certains cas un impact sur la qualité des données. Il convient donc d'avoir régulièrement un rapport mentionnant les défaillances qui ont eu lieu durant une période déterminée. Ceci permettra par

exemple d'associer des problèmes identifiés par les deux types de rapports précédents à une défaillance technique.

Afin de pouvoir tirer le meilleur parti de ces différents rapports, la définition de *benchmarks* est indispensable. Nous nous référons également à la techniques du ***data tagging*** présentée au chapitre 5 qui sera d'une aide précieuse pour établir ces rapports.

4ère Partie : Gestion journalière des données

9. Gérer la qualité des données : le *data management*

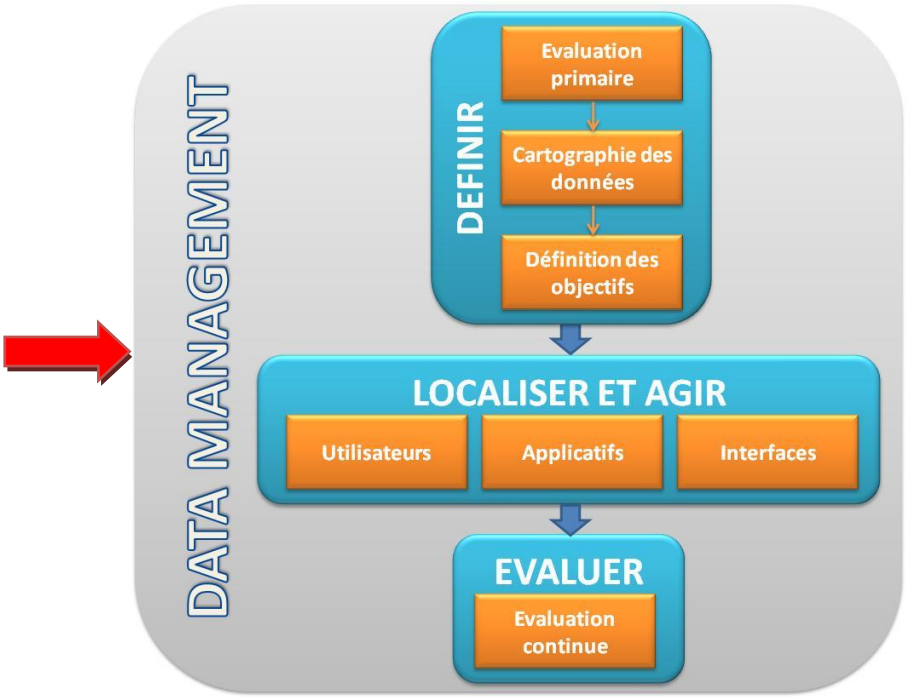


Figure 33 : Le data management a pour fondements les moyens et techniques de l'aide à la validation des données.

Ce sont principalement, comme nous l'avons mentionné dans nos hypothèses, les entreprises de grandes tailles qui sont les plus impactées par la problématique de la qualité de données. En effet, ce sont majoritairement celles-ci qui procèdent à la collecte massive de données. Celle-ci rend impossible le traitement manuel des données et implique donc l'utilisation de techniques d'aide à la validation de données. Le fait qu'il s'agisse d'entreprises de grande taille entraine une plus grande complexité dans la mise en place de tels projets car il ne s'agit généralement pas uniquement de mobiliser une ou deux personnes pour atteindre cet objectif mais bien des dizaines de personnes avec des profils souvent fort différents allant du programmeur à l'utilisateur final.

Nous avons constaté que lorsqu'on aborde le problème de la qualité des données celui-ci devient tentaculaire : quels cas d'utilisation, quelles erreurs, quelles dimensions, quelles règles et quels standards, comment les documenter, quelles techniques ou procédures utiliser pour quels problèmes? La qualité des données et leur validation nécessitent donc la mise en place d'une organisation capable de gérer et de coordonner les multiples aspects et intervenants qu'implique cette problématique. Nous présenterons dans ce chapitre la notion de *data management* ainsi qu'un canevas permettant d'aider les entreprises à définir ses missions.

Pour présenter ce qu'est le data management, nous retenons la définition suivante proposée par [Mark Mosley, 2007] :

> *Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.*

Nous complétons cette définition en ajoutant qu'une telle organisation doit être mise en place au sein des départements métiers et doit être en relation étroite avec les départements informatiques. En effet, ce sont les départements métiers qui doivent gérer les données dont ils sont, en définitive, les propriétaires mais surtout dont ils ont la connaissance. Au besoin, le data management demandera aux départements informatiques d'agir.

9.2 Son rôle

Afin de définir quel devait être le rôle du Data Management au sein d'une entreprise, nous avons pris comme point de départ deux axes de réflexion: d'une part le cycle de vie de la donnée et d'autre part les quatre phases que nous reconnaissons à l'activité de *data management*.

9.2.1 Les quatre phases du *data management*

Le *data management* est une activité qui a selon nous quatre phases principales. Celles-ci s'inscrivent dans un processus cyclique comme présenté à la figure 34. Les données changeant continuellement, elles doivent donc être contrôlées en permanence et leur qualité doit être régulièrement remise en question.

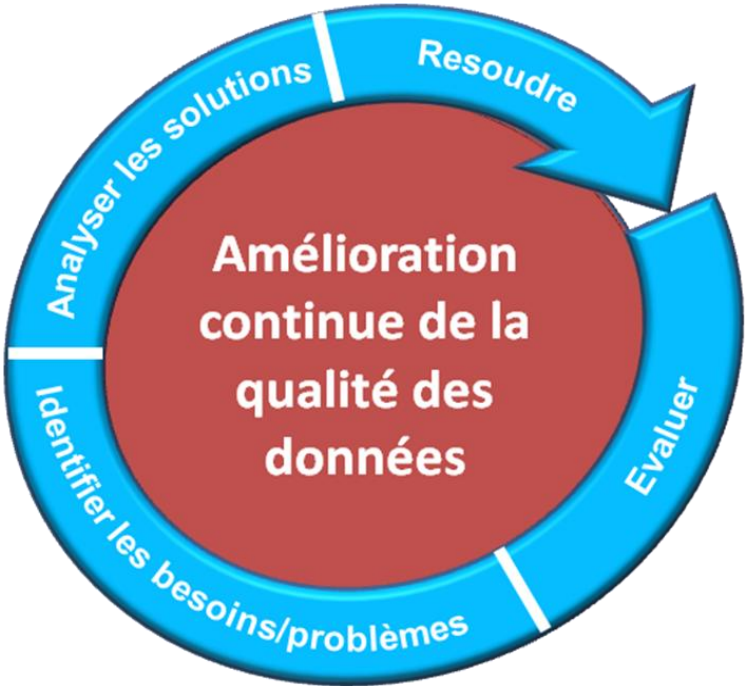


Figure 34: Les phases du data management

- Evaluer

  Cette phase consiste à évaluer continuellement la qualité des données comme nous l'avons décrit dans le chapitre 7, c'est-à-dire en adoptant un rôle de surveillance qui aura un caractère plus passif. Dans cette phase, le *data management* aura également un rôle plus actif voire proactif en évaluant les risques sur la qualité des données notamment en intervenant dès la conception de projets ayant un impact sur les données.

- Identifier

  Lorsque, durant la phase d'évaluation, une dégradation de la qualité des données ou un risque de dégradation est détecté, le

data management se doit de décrire clairement les origines du problème ou du risque. Il conviendra notamment de comprendre si l'origine du problème ou du risque est conjoncturelle (défaillance temporaire de l'applicatif…) ou structurelle (erreur récurrente dans l'encodage des adresses…). Cette phase identifiera alors la nécessité ou non de mettre en place des outils d'aide à la validation des données et au besoin, les objectifs de qualité à atteindre.

- Analyser

  Lorsque la décision de mettre en place des méthodes d'aide à la validation des données afin de solutionner un problème de qualité ou de diminuer un risque est prise, le *data management* doit procéder à une analyse des différentes solutions qui peuvent être mises en place pour permettre d'atteindre les objectifs fixés dans la phase précédente.

- Résoudre

  Cette dernière phase consiste à retenir une solution parmi celles proposées et à la mettre en place. Cette dernière terminée, le cycle recommence.

9.2.2 Le cycle de vie d'une donnée

Nous avons identifié quatre étapes dans le cycle de vie d'une donnée au sein d'une organisation : le projet, l'opérationnel, le *reporting*, et l'archivage.

Nous décrivons ces étapes comme suit :

- Le projet

  Cette étape reprend la définition des besoins en termes de données ainsi que les règles et les standards qui y seront associés dans le cadre d'un projet applicatif ou d'un projet purement organisationnel. Ce dernier cas doit aussi être abordé. Un changement au niveau des processus organisationnels manipulant des données peut également

avoir un impact sur celles-ci sans pour autant avoir nécessité de développement applicatif.

- L'opérationnel

  Lors de cette phase les données sont créées, lues, modifiées, voire supprimées par les processus opérationnels. Nous définissons donc cette phase comme l'utilisation des données dans un contexte opérationnelle.

- Le *reporting*

  Le *reporting* est une étape particulière où, par définition, les données ne sont pas modifiées mais agencées de telle manière à pouvoir créer de l'information. Dans cette phase, le *data management* a également un rôle crucial à jouer étant donné la complexité des flux de données qui alimentent les rapports des grandes entreprises (sources multiples, règles de calcul…). Cette complexité nécessite une gestion claire de ces flux.

- L'archivage

  L'aspect de l'archivage des données est souvent mal géré par les entreprises car il n'est que peu ou pas abordé dans la mise en place de nouvelles applications. Ceci est notamment dû à un manque de visibilité sur ce que seront les besoins en termes d'archivage dans le futur. C'est pour ces raisons que nous identifions cet aspect comme étant une étape du cycle de vie d'une donnée devant être gérée par le *data management*.

9.2.3 La gouvernance des données

La gouvernance des données (*data governance*) définit la politique globale de l'entreprise, en matière de gestion de données. En effet, aux côtés des quatre phases présentées, le *data management* doit également définir les standards et les règles qui devront être appliqués aux données dans les nouveaux projets de l'entreprise, s'assurer de la cohérence des mesures qui seront mises en place au sein de l'organisation et définir une stratégie pour la gestion des données.

9.2.4 La matrice des missions

Sur base des deux axes définis ci-dessus, nous avons créé une matrice permettant de faciliter la définition des missions que doit accomplir le *data management*. Une mission est une action entreprise par le *data management* à un moment précis du cycle de vie d'une donnée et dans le cadre d'une des quatre phases décrites auparavant.

Volontairement, nous n'avons pas souhaité proposer de missions prédéfinies. En effet, nous avons démontré tout au long de ce mémoire que la qualité des données dépend de son contexte d'utilisation, il en va de même pour la gestion des données. Dès lors, les missions seront spécifiques à chaque entreprise et fonction de l'importance accordée au *data management*.

La matrice proposée est représentée à la figure 35.

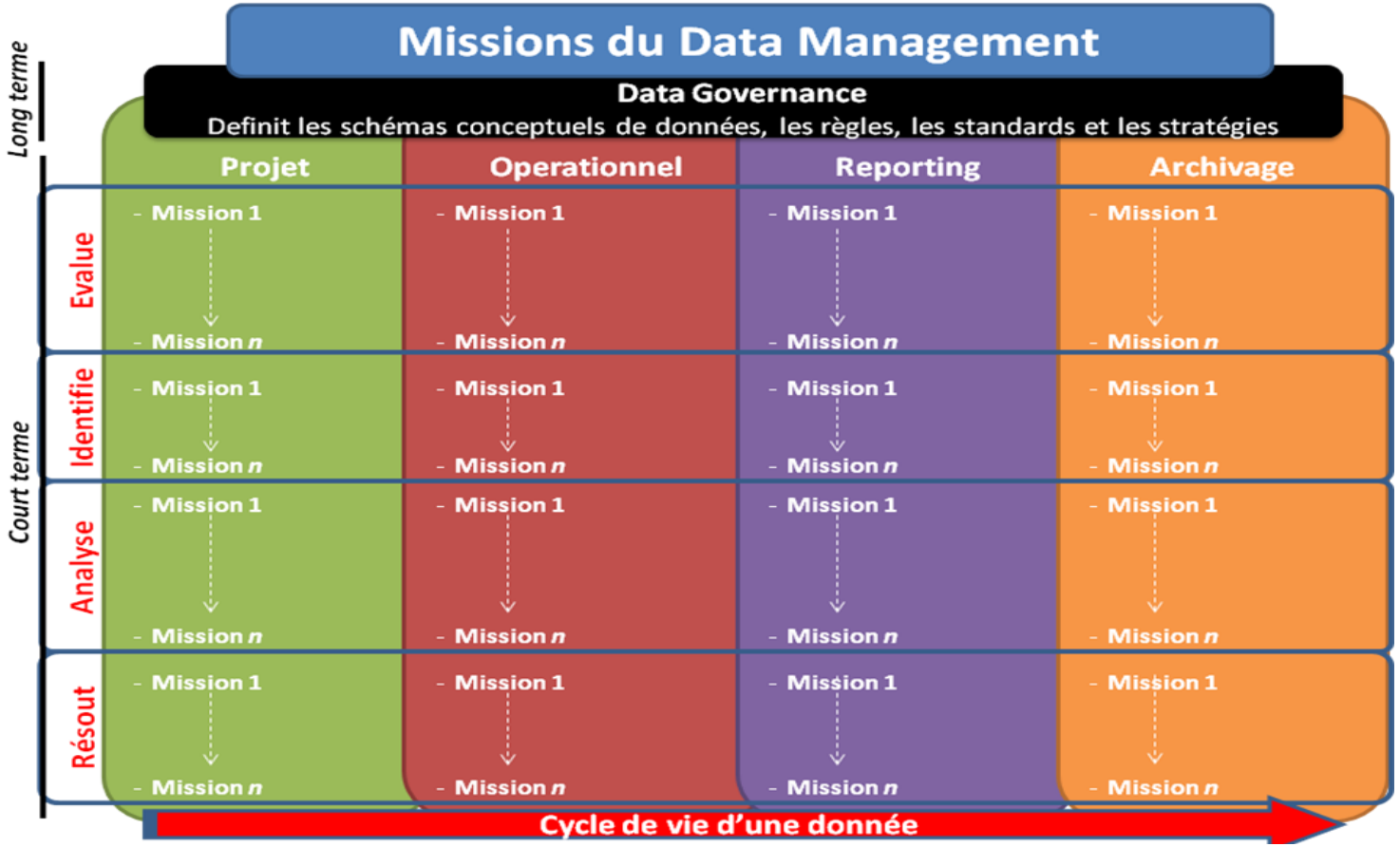


Figure 35 *: les missions du data management*

5ème Partie : Conclusion

Conclusion

L'objectif de ce mémoire était dans un premier temps de mettre en avant les différentes solutions informatiques disponibles pouvant faciliter la validation des données et permettre d'atteindre le niveau de qualité souhaité. Nous avons toujours considéré la qualité des données comme indissociable de l'aide à la validation des données.

Au fil de nos recherches et surtout au travers de nos rencontres professionnelles, nous nous sommes vite rendu compte que la complexité de l'aide à la validation des données ne se trouvait pas au niveau des solutions techniques mais bien dans l'organisation de ces solutions et dans la définition des règles de validation par les départements métiers. Nous nous sommes également aperçu que si le problème de la qualité des données était connu de tous, il était considéré comme un problème purement informatique par les départements métiers. C'est principalement sur base de ce constat et en opposition avec celui-ci que nous avons décidé d'aborder l'aide à la validation des données d'un point de vue *business* plutôt que purement informatique. C'est cette approche *business* de l'aide à la validation données qui est l'apport principal de ce mémoire. Nous sommes en effet convaincus que si l'informatique permet de faciliter la validation des données, c'est avant tout aux départements métiers de prendre l'initiative et de la conduire. Nous avons voulu remettre la responsabilité de l'utilisateur au centre du processus de validation.

Si la littérature est abondante sur les sujets de la qualité des données, du *data management*, des outils de *data cleaning*, ou encore du *data profiling*, ils sont souvent abordés de manière isolée sous un angle très (trop) technique ou à contrario sous un angle très (trop) généraliste. Il est donc difficile pour une entreprise qui souhaite améliorer la qualité de ses données d'avoir une vision concrète des tenants et aboutissants d'une telle activité. Nous avons tenté de positionner ce mémoire comme un **guide** pour la mise en place d'une organisation, au sein des départements métiers, capable de gérer et de faciliter la validation des données et d'ainsi en accroître leur qualité. Nous avons en permanence tenté de rapprocher les aspects théoriques du sujet étudié à des cas concrets tout en proposant des solutions pratiques.
Ainsi, dans la première partie, après avoir défini ce qu'était la qualité des données, nous avons voulu sensibiliser le lecteur aux conséquences que

pouvait engendrer une mauvaise qualité tout en précisant que le but ne doit pas être, sauf exception, le zéro défaut.

Dans la seconde partie, nous avons présenté de manière concrète ce que peuvent être les données erronées et leurs différents types. Ensuite, nous avons donné un aperçu des solutions informatiques qui pouvaient être mises en place en portant une attention toute particulière à la documentation et aux méthodologies associées. En effet, nous considérons cet aspect comme primordial dans un processus d'amélioration de la qualité des données.

La troisième et la quatrième parties constituent le cœur de ce mémoire. Nous avons proposé dans ces deux parties une approche pour mettre en place une organisation pour la gestion des données au sein d'une grande entreprise. Celle-ci consiste dans un premier temps à faire une évaluation du besoin ou non d'une telle entité. Dans un second temps, sur base du contexte d'utilisation des données et de la documentation présentée dans la deuxième partie, nous avons proposé une méthodologie pour définir les objectifs généraux en termes de qualité. Nous avons ensuite envisagé les solutions purement métiers et préventives qui pouvaient être mises en place. Nous avons également présenté quelques techniques permettant d'évaluer le niveau de qualité obtenu. Enfin, nous avons proposé une matrice permettant de définir et de mettre en place une organisation visant à améliorer la qualité des données au sein d'une entreprise.

Certes, de nombreuses questions restent ouvertes. Entre autres, il conviendra de déterminer où se situe la frontière entre les responsabilités des départements informatiques et celles des départements métiers. Les rôles de chacun doivent être décrits et les interactions que ces départements ont dans le cadre de la gestion des données doivent faire l'objet d'études approfondies. Nous laissons également le soin à des travaux futurs de procéder à la validation du modèle d'évaluation de la criticité d'un attribut. Enfin, le rôle des progiciels et leur impact sur la connaissance et la maîtrise qu'a une entreprise sur ses données doivent également être approfondis. En effet, ce type d'outils pose souvent un voile opaque sur les données et les modèles de données clef sur porte qu'ils proposent ne nécessitent plus de réflexion sur les données de la part de l'entreprise avec le risque d'une perte de compétences des départements informatiques, notamment en termes de modélisation.

Pour conclure, ce mémoire n'avait pas la prétention d'analyser en détail tous les aspects abordés mais bien d'offrir une vue générale des enjeux liés à la validation des données tout en proposant une approche concrète du sujet dans le cadre des grandes entreprises. Le but, in fine, était d'ouvrir le débat

de la qualité des données à des aspects autres que purement techniques espérant ainsi sensibiliser le lecteur non informaticien.

Bibliographie

Livres

- Redman, T.C. (2000). Data Quality : The Field Guide. *Digital Press*.
- David Loshin (2001). Enterprise Knowledge Management, The Data Quality Approach. *Morgan Kaufmann*.
- Guy Marion (1997). Les systèmes d'information de gestion. *Edition De Boeck*
- J-L Hainaut (2002). Bases de données et modèles de calcul. 3ème édition. *Dunod.*
- Jack E. Olson (2003). Data quality-The accuracy dimension. *Morgan Kaufmann*.
- Olivier Heurtel (2007). PHP 5.2 – Développer un site Web dynamique et interactif. *Editions ENI*

Publications

- Boydens, I. (1998, janvier). Evaluer et améliorer la qualité des bases de données. *Techno 7- N°7* .
- Redman, T. C. (2004, août). Data: An Unfolding Quality Disaster. *Information Management Magazine*.
- Redman, T. C. (1995). Improve Data Quality for Competitive Advantage. Sloan Management Review 36, No 2, p. 99-107
- Yair Wand et Richard Y. Wang (1996, novembre). Anchoring Data Quality Dimensions in Ontological Foundations. *Comunication to the ACM-Vol.39, No. 11*.
- Helfert Markus (2001). Managing and measuring data quality in data warehousing. *Institute of Information Management, University of St. Gallen.*
- Erhard Rahm et Hong Hai Do (n.a). Data Cleaning: Problems and Current Approaches. *University of Leipzig, Germany*
- Peter Buneman, Sanjeev Khanna and Wang-Chiew Tan (n.a.). Data Provenance: Some Basic Issues. *University of Pennsylvania*
- Jonathan G. Geiger, Intelligent Solutions, Inc., Boulder, CO (n.a.). Data Quality Management: The Most Critical Initiative You Can Implement. *SUGI 29 Data Warehousing, Management and Quality - Paper 098-29*

- Dmitri V. Kalashnikov et Sharad Mehrotra (2005) Exploiting relationships for domain-independent data cleaning. *Computer Science Department, University of California*
- Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim et Doheon Lee (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery, 7, 81–99, 2003 Kluwer Academic Publishers*
- Andrew White, David Newman, Debra Logan, John Radcliffe (2006). Mastering Master Data Management. *GARTNER, ID Number: G00136958*
- Jean-Roch Meurisse (2004). Extraction de données de sites web: Méthodologie, outils et étude de cas. *Institut d'informatique, FUNDP*
- Leo L. Pipino, Yang W. Lee, et Richard Y. Wang (2002). Data Quality Assessment. *Communications of the ACM Vol. 45, No. 4.*
- Robert Rich (2008). Diagnosing Customer Data Disorder. *DM review*
- Mark Mosley (2007). DAMA-DMBOK Guide™ (Data Management Body of Knowledge). *DAMA International Foundation*
- Christine Parent et Stefano Spaccapietra (n.a.). Issues and Approaches of Database Integration. *Swiss Federal Institute of Technology*
- Mong Li Lee, Hongjun Lu, Tok Wang Ling et Yee Teng Ko (n.a). Cleansing Data for Mining and Warehousing. *School of Computing National University of Singapore*

Cours

- J-L. Hainaut,Ingénierie des bases de données, 2003

Sites internet

- Communauté Wikipédia. (n.a.). *Gestion de la relation client*. Consulté le 30 mai 2009, sur Wikipédia: http://fr.wikipedia.org/wiki/ Gestion_de_la_relation_client
- Communauté Wikipédia. (n.a.). *Syntaxe*. Consulté le 22 mai 2009, sur Wikipédia: http://fr. wikipedia.org/wiki/Syntaxe
- Communauté Wikipédia. (n.a.). *Data flow diagram*. Consulté le 20 juillet 2009, sur Wikipédia : http://en.wikipedia.org/wiki/ Data_flow _diagram

- Diffen (n.a.). *Data vs Information - Differences in meaning*. Consulté le 18 avril 2009 sur http://www.diffen.com/difference/Data_vs_Information
- SystemeETL. (n.a.). Description des composantes d'un système ETL. Consulté le 5 mai 2009, sur http://www.systemeetl.com/back_front_4.htm

Conférences

- Karel-Forrester, R. (2008). Ensuring The Value Of Your Trusted Data. *Data Quality Summit '08*. Eindhoven.
- Cappiello Cinzia (2009). Quantifying the Strategic Value and Cost of Data Quality Improvement. *Marcus Evans: 2nd Annual Data Quality Management in the Energy Sector*. Barcelona.
- Carmen Artigas - Synergic Partners (2009). Beyond Data Quality: The Path to Data Governance. *Marcus Evans: 2nd Annual Data Quality Management in the Energy Sector*. Barcelona.